УНИВЕРЗИТЕТ У НОВОМ САДУ

ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА У НОВОМ САДУ

Никола Николић

Аутоматско издвајање мишљења из текстуалних коментара студентских анкета

ДОКТОРСКА ДИСЕРТАЦИЈА

Ментор:

проф. др Александар Ковачевић

Нови Сад, 2021.

Редни број, **РБР**:	
Идентификациони број, **ИБР**:	
Тип документације, **ТД**:	Монографска публикација
Тип записа, **ТЗ**:	Текстуални штампани материјал
Врста рада, **ВР**:	Докторска дисертација
Аутор, **АУ**:	Никола Николић
Ментор, **МН**:	др Александар Ковачевић, ванредни професор
Наслов рада, **НР**:	Аутоматско издвајање мишљења из текстуалних коментара студентских анкета
Језик публикације, **ЈП**:	Српски
Језик извода, **ЈИ**:	Српски / Енглески
Земља публиковања, **ЗП**:	Република Србија
Уже географско подручје, **УГП**:	Аутономна Покрајна Војводина
Година, **ГО**:	2021
Издавач, **ИЗ**:	Факултет техничких наука
Место и адреса, **МА**:	Нови Сад, Трг Доситеја Обрадовића 6
Физички опис рада, **ФО**: (поглавља/страна/ цитата/табела/слика/графика/прилога)	(8/218/106/28/66/0/1)
Научна област, **НО**:	Електротехничко и рачунарско инжењерство
Научна дисциплина, **НД**:	Аутоматска обрада природног језика
Предметна одредница/Кqучне речи, **ПО**:	аутоматска обрада природног језика, аспектно базирана сентимент анализа, анализа мишљења, анализа текста, анализа података, коментари, студентске анкете
УДК	
Чува се, **ЧУ**:	Библиотека Факултета техничких наука, Трг Доситеја Обрадовића 6, 21000 Нови Сад
Важна напомена, **ВН**:	

Извод, **ИЗ**:		Регрутовање нових и задржавање постојећих студената су важна питања за све високошколске установе. Стога је пресудно стално праћење нивоа задовољства студената. Аутоматска анализа мишљења студената се може реализовати применом аспектно базиране сентимент анализе (АБСА). АБСА је под-дисциплина обраде природног језика која се фокусира на идентификацију сентимената (негативних, неутралних, позитивних) и аспеката (носиоца сентимента) у реченици. Циљ ове докторске дисертације је да предложи систем за АБСА текстуалних коментара студентских анкета на српском језику. Предложени систем се ослања на технике обраде природног језика, модела машинког учења, правила и речника. Корпус је прикупљен и анотиран за развој и евалуацију система и укључује рецензије студената о наставном особљу и студијским програмима на Факултету техничких наука. Резултати истраживања показују да се позитивни сентимент може успешно идентификовати са Ф-мером 0,91, док се негативан сентимент може идентификовати са Ф-мером 0,97. Док су Ф-мере за аспекте у опсегу између 0,49 и 0,89, у зависности од њихове учесталости у корпусу. Према сазнању аутора, ово је прво истраживање АБСА које је спроведено на нивоу сегмента реченице за српски језик. Методологија и сазнања која су представљена у овој докторској дисертацији пружају преко потребне основе за даљи рад на анализи сентимената за српски језик који је у овој области недовољно истражен и има недостатак језичких ресурса.	
Датум прихватања теме, **ДП**:		31.10.2019.	
Датум одбране, **ДО**:			
Чланови комисије, **КО**:	Председник:	др Драган Ивановић, редовни професор Факултет техничких наука, Нови Сад	
	Члан:	др Бојана Димић Сурла, редовни професор Рачунарски факултет, Београд	
	Члан:	др Александар Купусинац, ванредни професор Факултет техничких наука, Нови Сад	Потпис ментора
	Члан:	др Јелена Сливка, ванредни професор Факултет техничких наука, Нови Сад	
	Члан, ментор:	др Александар Ковачевић, ванредни професор Факултет техничких наука, Нови Сад	

Accession number, **ANO**:	
Identification number, **INO**:	
Document type, **DT**:	Monographic publication
Type of record, **TR**:	Textual printed material
Contents code, **CC**:	Ph.D. thesis
Author, **AU**:	Nikola Nikolic
Mentor, **MN**:	Aleksandar Kovacevic, Ph. D., Associate professor
Title, **TI**:	Automatic opinion extraction from textual comments of students surveys
Language of text, **LT**:	Serbian
Language of abstract, **LA**:	Serbian / English
Country of publication, **CP**:	Serbia
Locality of publication, **LP**:	Autonomous Province of Vojvodina
Publication year, **PY**:	2021
Publisher, **PB**:	Faculty of Technical Sciences
Publication place, **PP**:	Novi Sad, Trg Dositeja Obradovića 6
Physical description, **PD**: (chapters/pages/ref./tables/pictures/graphs/appendixes)	(8/218/106/28/66/0/1)
Scientific field, **SF**:	Electrical and computer engineering
Scientific discipline, **SD**:	Natural language processing
Subject/Key words, **S/KW**:	natural language processing, aspect-based sentiment analysis, opinion analysis, text analysis, data analysis, comments, student surveys
UC	
Holding data, **HD**:	Library of Faculty of Technical Sciences, Trg Dositeja Obradovića, 21000 Novi Sad
Note, **N**:	

Abstract, **AB**:		Student recruitment and retention are an important issue for all higher education institutions. Constant monitoring of student satisfaction levels is therefore crucial. Aspect-based sentiment analysis is a sub-discipline of natural language processing (NLP) that focuses on the identification of sentiments (negative, neutral, positive) and aspects (sentiment targets) in a sentence. This research introduces a system for aspect-based sentiment analysis of free text reviews expressed in student opinion surveys in the Serbian language. Sentiment analysis was carried out at the finest level of text granularity - the level of sentence segment (phrase and clause). The presented system relies on NLP techniques, machine learning models, rules, and dictionaries. The corpora collected and annotated for system development, and evaluation comprise: students' reviews of teaching staff at the Faculty of Technical Sciences. The research results indicate that positive sentiment can successfully be identified with F-measure of 0.91 while negative sentiment can be detected with F-measure of 0.97. While the F-measures for the aspects are in range from 0.49 to 0.89, depending on their frequency in the corpus. To the best of the authors' knowledge, this is the first study of aspect-based sentiment analysis carried out at the level of the sentence segment for the Serbian language. The methodology and findings presented in this paper provide a much-needed bases for further work on sentiment analysis for the Serbian language that is well under-resourced and under-researched in this area.	
Accepted by the Scientific Board on, **ASB**:		31.10.2019.	
Defended on, **DE**:			
Defended Board, **DB**:	President:	Dragan Ivanović, Ph.D., full prof. Faculty of Technical Sciences, Novi Sad	
	Member:	Bojana Dimić Surla, Ph.D., full prof. Faculty of Computer Science, Beograd	
	Member:	Aleksandar Kupusinac, Ph.D., assoc. prof. Faculty of Technical Sciences, Novi Sad	Menthor's sign
	Member:	Jelena Slivka, Ph.D., assoc. prof. Faculty of Technical Sciences, Novi Sad	
	Member, Mentor:	Aleksandar Kovačević, Ph.D., assoc. prof. Faculty of Technical Sciences, Novi Sad	

Захвалница

Велику захвалност дугујем свом ментору, проф. др Александру Ковачевићу, на огромној подршци и несебичној помоћи пруженој приликом израде ове докторске дисертације. Захваљујући његовим научним, стручним и педагошким саветима, ова докторска дисертација је доведена до краја.

Захваљујем се члановима Комисије на драгоценим саветима и корисним сугестијама приликом израде ове докторске дисертације.

Највећу захвалност дугујем својој породици на разумевању и подршци.

Резиме

Повећање броја интернет страница и платформи је утицало на пораст броја текстуалних података о исказаном мишљењу у дигиталном облику. Људи су почели да се повезују једни са другима и да на слободан начин износе своја мишљења и ставове о било којој теми без обзира на њихову географску локацију. Стога се од 2002 године повећао број истраживања у области сентимент анализе са циљем издвајања корисних информација попут аспекта о ком људи причају и која су њихова мишљења. Последњих година је примена анализе мишљења и сентимента проширена на скоро све домене, од потрошачких производа, мултимедије, финансија, социјалних догађаја, политике итд. Добијене информације се могу искористити у апликацијама попут система за преоруку производа и сервиса или система за подршку у одлучивању.

У овој докторској дисертацији је обрађена област анализе мишљења у текстуалним коментарима на српском језику. Аутоматско издвајање мишљења, или сентимент анализа, је под-област области обраде природног језика која се бави откривањем заједничких образаца (шаблона, карактеристика) мишљења људи из неструктуираног садржаја (слободног текста). Конкретније, аутоматско издвајање мишљења из текста подразумева издвајање става или осећаја (означеном као сентимент) који је особа испољила ка одређеном објекту (означеном као аспект). Сентимент се најчешће дели на три типа: позитиван, негативан и неутралан. Аутоматско издвајање мишљења може бити вршено на различитим текстуалним целинама (нивоима) попут документа, реченице, и сегмента реченице.

Главни циљ овог истраживања је формирање модела за аутоматско издвајање мишљења студената исказаних у текстуалним коментарима на српском језику. У разматрање су узети коментари студената исказаних путем обавезних анкета које су спроведене на Факултету Техничких Наука, Универзитета у Новом Саду. На Факултету Техничких Наука, анкетни процес се спроводи кроз пет врста анкета за преко 14.000 студената. Свака анкета пружа могућност студентима да путем текстуалних коментара у слободној форми изразе своје мишљење (задовољство или незадовољство) о једном или више аспеката студирања. Ручна обрада и анализа великог броја текстуалних коментара је временски захтевна и подложна грешкама људског фактора. Стога је јасна потреба за моделом система који ће вршити аутоматско издвајање мишљења из коментара.

Предложена методологија је базирана на коришћењу више типова модела - модела заснованим на речницима и правилима, модела машинског учења, где се убрајају и модели дубоког учења. Модели заснованим на правилима и речницима су се показали да могу омогућити значајно побољшање перформанси само за корпус на основу ког су и развијени. Интеграција модела заснованом на правилима и речницима са класичним моделима машинског учења може бити од користи без обзира на корпус. Вишејезични модели дубоког учења са применом трансфера знања су се показали веома ефикасним у задатку идентификације сентимента. У случају идентификације аспеката, модели дубоког учења су показали сличну успешност као и класични модели машинског учења. За побољшање перформанси овог модела је потребна већа количина анотираног корпуса за све класе аспеката којим би се квалитетније дообучавали модели дубоког учења.

Експериментални резултати показују да аспектно базирана сентимент анализа за домен високог образовања за српски језик може успешно применити на нивоу сегмента реченице. Резултати идентификације сентимента у коментарима студената показују да се позитиван сентимент може успешно идентификовати са Ф-мером од 0,91, док се негативан сентимент може идентификовати са Ф-мером од 0,97. Ниво успешности идентификације аспеката варира у распону између 0,49 и 0,89, у зависности од различитих фактора попут квалитета анотација у обучавајућем скупу, лексичке варијабилности и количине анотираних података.

Приказани приступ анализе мишљења има потенцијал широке примене у свим високошколским установама у Србији где је неопходна аутоматска анализа велике количине рецензија из студентских анкета. То подразумева да се у високо образовним установама у Србији може изводити истраживање јавног мњења великих размера. Аутоматизација процеса анализе мишљења студената ће умањити време и труд који су потребни људским кустосима и омогућити управи факултета праћење задовољства, односно незадовољства, одређеним аспекатима студирања. Такође, резултати овог истраживања могу побољшати квалитет сервиса које пружају високошколске установе и помоћ у регрутовању нових и задржавању постојећих студената.

Abstract

The increase in the number of websites and platforms has led to an increase in the number of textual opinion data expressed in digital form. People began to connect with each other and to freely express their opinions and views on any topic, regardless of their geographical location. Therefore, since 2002, the number of researches in the field of sentiment analysis has increased with the aim of extracting useful information such as the aspect that people talk about and what their opinions are. In recent years, the application of opinion and sentiment analysis has expanded to almost all domains, from consumer products, multimedia, finance, social events, politics, etc. The information obtained can be used in applications such as product and service recommendation systems or decision support systems.

The topic of this doctoral dissertation is automatic opinion analysis of textual comments in the Serbian language. Automatic opinion extraction, or sentiment analysis, is a sub-discipline of the field of natural language processing that deals with the discovery of common patterns (characteristics) of people's opinions from unstructured content (raw text). More specifically, the automatic extraction of opinion from the text implies the extraction of an attitude or feeling (denoted as sentiment) that a person has expressed towards a certain object (denoted as an aspect). Sentiment is usually classified as positive, negative and neutral. Automatic extraction of opinions can be performed at different levels such as a document, a sentence, and a sentence segment.

The main goal of this research is to develop a model for the automatic extraction of students' opinions expressed in textual comments in the Serbian language. The comments of students expressed through mandatory surveys conducted at the Faculty of Technical Sciences, University of Novi Sad were taken into consideration. The survey process at the Faculty of Technical Sciences includes five types of surveys for over 14,000 students annually. Each survey provides an opportunity for students to through textual comments express their opinion on one or more aspects regarding various aspects of their student experience. Manual processing and analysis of a large number of textual comments is time consuming and prone to human error. Therefore, there is a clear need for a system that will automatically extract opinions from comments.

The proposed methodology is based on the use of several types of models - models based on dictionaries and rules, machine learning models, which include deep learning models. Rule-based and dictionary-based models have been shown to be able to significantly improve performance only for the corpus on which they were developed. Integrating rule-based and vocabulary-based models with common machine learning models can be beneficial regardless of corpus. Multilingual deep learning models with the application of transfer learning have proven to be very effective in the task of identifying sentiment. In the case of aspect identification, deep learning models have shown similar success as common machine learning models. To improve the performance of this model, a larger amount of annotated corpus is needed for all classes of aspects, which would provide better training for deep learning models.

The experimental results of this doctoral dissertation show that an aspect-based sentiment analysis in the domain of higher education for the Serbian language can be successfully applied at the sentence segment level. The results of sentiment identification in student comments show that positive sentiment can be successfully identified with an F-measure of 0.91, while negative sentiment can be identified with an F-measure of 0.97. Performance of the identifying aspects varies between 0.49 and 0.89, depending on various factors such as the quality of the annotations in the training set, lexical variability and the amount of annotated data.

The presented approach of opinion analysis has the potential of wide application in all higher education institutions in Serbia where automatic analysis of a large number of reviews from student surveys is necessary. This means that large-scale public opinion polls from student surveys can be conducted. Automating the process of analyzing student opinions will reduce the time and effort required by human curators and enable faculty management to monitor student satisfaction with certain aspects of studying. Also, the results of this research can improve the quality of services provided by higher education institutions and help in recruiting new and retaining existing students.

Садржај

Списак слика

Списак Табела

1 Увод и мотивација

Образоване установе широм света троше велике своте новца како би побољшали своје услове и услуге, задржали постојеће студенте и уписали нове. Према истраживању које је спроведено 2020. године у Америци (Levitz, 2020), четворогодишњи државни универзитет за упис једног студента потроши у просеку 470 долара, док приватни универзитет и преко четири пута више. Највећи проценат потрошеног новца се издвајао у дигитални маркетинг, као и за друге пријемне догађаје у уписној кампањи. Аутор је упоређивањем трошкова образованих установа током уписа установио да поменути трошкови представљају кључно мерило повратка инвестиције тј. уложеног новца. Уколико однос уложеног новца и остварених резултата током уписа значајно одступа од просека приказаног у истраживању, аутор закључује да вероватно постоје проблеми везани за стратегију уписа, ефикасност примене стратегије и продор на тржишту. Из наведеног истраживања се може закључити да је за успех образоване установе веома значајна инвестиција у маркетинг током уписне кампање.

У другом истраживању (Rauschnabel *et al.,* 2016) је анализиран утицај мишљења и задовољства студената о квалитету услуга образоване установе на углед и репутацију образоване установе. Аутори су истакли да је мишљење и задовољство студената веома значајно и да представља једну врсту маркетинга, познатију као маркетинг „од уста до уста“ (енгл. *word-of-mouth - WoM*) (Li, 2013) и маркетинг „од уста до уста“ електронским путем (енгл. *electronic word-of-mouth - eWoM*) (Yahya *et al.,* 2014). Резултати истраживања показују да задовољство студената и позитиван маркетинг од уста до уста веома утиче на бренд и репутацију факултета, самим тим и на упис студената.

Студенти своје мишљење изражавају кроз редовне анкете и упитнике које се спроводе у образованим установама и путем друштвених мрежа и форума на интернету. Према важећем закону о високом образовању Републике Србије евалуација студија и педагошког рада наставника је обавезна и спроводи се у свим образованим установама. Основни циљ евалуације јесте утврђивање квалитета педагошког рада наставника, студијских програма и функционисања служби образоване установе, у циљу корекције

наставног процеса и рада просветних радника ради подизања квалитета. Са друге стране, коментари постављени путем друштвених мрежа и форума на интернету такође нису занемарљиви. На тај начин се врши *eWoM* маркетинг који утиче на репутацију образоване установе и уписни циклус.

Већи део званичних анкета и упитника у образованим установама је у структуираном облику које пружају могућност да се напише коментар у облику слободног текста, односно неструктуираног садржаја. Ови коментари, заједно са коментарима са интернета, носе корисне информације које је потребно анализирати. Неструктуирани коментари се типично индексирају и претражују помоћу кључних термина и класичних техника проналажења информација (енгл. *Information Retrieval - IR*). Међутим, претрагом помоћу кључних термина није могуће на једноставан начин пронаћи коментаре у којима су студенти исказали одређени сентимент поларитет (позитиван, негативан, неутралан) према одређеном аспекту студирања (наставник, организација предмета, настава итд.). Индексирање неструктуираних коментара семантичким подацима попут аспекта и сентимент поларитета би омогућило ефикаснију претрагу, анализу и сумирање мишљења студента. На тај начин би управа образоване установе могла на једноставан начин да идентификује којим аспектом образоване установе студенти нису задовољни и за које ће индиректно вршити негативан маркетинг (*WoM*, *eWoM*). Међутим, за индексирање коментара аспектом и сентимент поларитетом потребно је анотирати (означити) одговарајуће делове докумената који се односе на циљане аспекте и уз то им доделити сентимент поларитет. Ручна анотација сентимента и аспекта је због велике количине докумената временски захтевна и непрактична. Из тог разлога је потребна аутоматска обрада коментара и анотација семантичким подацима. Процес аутоматске обраде тј. издвајања сентимента и аспеката из неструктуираног садржаја се назива аспектно базирана сентимент анализа (енгл. *Aspect-Based Sentiment Analysis - ABSA*) која је поддисциплина области анализе сентимента, односно мишљења (енгл. *opinion mining*) (Liu, 2015; Thet *et al.,* 2010). Циљ анализе мишљења је откривање заједничких образаца (шаблона, карактеристика) мишљења корисника у циљу унапређења пословања, разумевања корисничких избора, њихових намера и осећаја у реалном времену (Liu, 2015).

Иако постоји велики број система и методологија за анализу сентимента они су зависни од домена и језика, развијених углавном за водеће језике попут енглеског језика. Већина *ABSA* система врши анализу на нивоу реченица али то није довољно за прецизније издвајање информација. Разлог тога је случај у коме се у једној реченици налази више аспеката и сентимент поларитета. На пример, коментар „Професор је добар педагог, али је предмет организован веома лоше“, јасно указује на то да се у првом делу реченице помиње наставник у позитивном контексту док се у другом делу реченице предмет помиње у негативном контексту. Стога је неопходан *ABSA* систем за

анализу студентских коментара који је прилагођен домену високог образовања и који ће бити у стању да идентификује више аспеката и сентимент поларитета у реченици, узимајући у обзир језик који се користи у коментару.

У овој докторској дисертацији је описан систем за аспектно базирану сентимент анализу на нивоу сегмента реченице (клаузе или фразе) који је прилагођен за домен високог образовања у српском језику. Систем се састоји из неколико делова. Прво се коментар дели на реченице и онда на фразе и клаузе, коришћењем разделника развијеног у сврху овог истраживања. Сваки сегмент реченице се затим анотира (означава) са једним од предефинисаних аспеката помоћу модула који се састоји од више модела машинског учења (енгл. *Machine Learning - ML*). Затим се анотираним аспектима додељује један од два сентимент поларитета (позитивни, негативни)[1] на основу одвојених модула заснованим на *ML*. Систем је евалуиран на два засебно анотирана корпуса. Први корпус је сакупљен из званичних студентских анкета са Факултета техничких наука, Универзитета у Новом Саду. Други корпус је сачињен од коментара са интернет сајта „Оцени професора“[2] на основу ког је извршена компаративна анализа. Шест аспеката (Наставник, Настава, Однос са студентима, Предмет, Материјали и Организација) је успешно идентификовано са просечном Ф-мером од 81 посто, док је сентимент поларитет идентификован са просечном Ф-мером од 96 посто. Постигнути резултати указују на то да се предложени систем може користити за формирање *IR* система који омогућава семантичку претрагу и анализу засновану на сентимент поларитету и / или аспектима. Такав тип *IR* система може побољшати квалитет сервиса и услуга у високошколским установама.

Досадашња *ABSA* истраживања у домену високог образовања су базирала своју анализу само на нивоу докумената и реченица. Истраживање које је приказано овом докторском дисертацијом представља допринос истраживању у *ABSA* области у домену високог образовања тиме што је идентификација аспеката и додељивање сентимент поларитета извршено на финијем нивоу прецизности – нивоу сегмента реченице. Експериментални резултати показују да се *ABSA* може успешно извести на нивоу сегмента реченице, омогућавајући аутоматско издвајање најдетаљнијих информација о мишљењу студената из анкета. Користећи ове информације, високошколске установе могу побољшати квалитет својих услуга и квалитет процеса доношења одлука. Методологија и ресурси који су представљени у овом раду доприносе области обради природног језика (енгл. *Natural Language Processing - NLP*) на српском језику која је недовољно истражена, посебно у *ABSA* поддисциплини. По сазнању аутора ове

[1] Аутоматска анализа сентимента неутралног поларитета није разматрана у овом истраживању због његове изузетно мале учесталости у корпусу

[2] Веб сајт за оцену професора у Србији: http://oceniprofesora.com/

докторске дисертације не постоји јавно доступни системи за аутоматско издвајање сентимента и аспеката на нивоу сегмента текста (фраза и клауза) за српски језик, прилагођен за домен високог образовања.

У наставку је описан приступ истраживању почевши од предмета истраживања, циља истраживања и постављених хипотеза. На самом крају поглавља је наведена структура дисертације по поглављима која су описана.

1.1 Предмет истраживања

На основу уводног излагања и мотивације је дефинисан предмет истраживања докторске дисертације. Главни предмет овог истраживања јесте формирање модела за аутоматско издвајање мишљења студената исказаних у текстуалним коментарима. У разматрање се узимају коментари на српском језику исказани путем обавезних анкета које су спроведене на Факултету Техничких Наука (ФТН), Универзитета у Новом Саду. На Факултету Техничких Наука, анкетни процес се спроводи кроз пет врста анкета за преко 14.000 студената. Свака анкета пружа могућност студентима да путем текстуалних коментара у слободној форми изразе своје мишљење (задовољство или незадовољство) о једном или више аспеката студирања. Ручна обрада и анализа великог броја текстуалних коментара је временски захтевна и подложна грешкама људског фактора. Стога је јасна потреба за моделом система који ће вршити аутоматско издвајање мишљења из коментара који ће управи факултета омогућити мониторинг одређених аспеката студирања (нпр. наставник, предмет, материјали, итд.), као и детектовање задовољства, односно незадовољства, истим. Додатно, за евалуацију модела система су коришћени и коментари са интернета из недавног истраживања сентимента у српском језику (Grljević, 2016).

Прегледом релевантне литературе утврђено је да модели за аутоматско издвајање мишљења зависе од језика за који су развијени. Највећи број модела развијен је за енглески језик. Поред тога утврђено је да модели зависе од домена примене, односно да се нпр. модели развијени за домен рецензија производа не могу без измена успешно применити на домен рецензија филмова итд. Такође, већина постојећих модела омогућава издвајање мишљења само на нивоу документа што за случај коментара са студентских анкета није довољно детаљно јер коментар може садржати више реченица од којих свака може садржати више различитих аспеката и сентимента.

Из горенаведеног проистиче предмет (проблем) истраживања тезе који представља формирање модела за аутоматско издвајање мишљења из коментара студентских анкета који има следеће карактеристике:

- омогућава обраду коментара на српском језику,
- прилагођен је домену студентских анкета,
- омогућава издвајање мишљења на најдетаљнијем нивоу (сегменту реченице).

По најбољем сазнању аутора не постоји модел за аутоматско издвајање мишљења на нивоу сегмента реченице за српски језик, прилагођен за домен студентских анкета.

1.2 Циљ истраживања и хипотезе

Основни циљ овог истраживања је развијање модела система за аутоматско издвајање мишљења из коментара студентских анкета који ће решити дефинисани проблем овог истраживања. Дакле, може се дефинисати следећа претпоставка тј. хипотеза:

- **Хипотеза**: Могуће је развити модел за аутоматску обраду коментара на српском језику, прилагођеном домену високог образовања, тако да омогућава издвајање мишљења на нивоу сегмента реченице.

Сходно основном циљу и постављеној хипотези, може се идентификовати више потциљева, односно корака ка постизању основног циља и постављених хипотеза тј. претпоставки. Један од првих корака јесте проучавање теоријских основа постојећих система за анализу мишљења за водеће језике и српски језик. Акценат проучавања теоријских основа се ставља на проучавање актуелног стања у области аутоматског издвајања мишљења. Наредни корак је проучавање доступних алата и ресурса обраде природног језика на нивоу сегмента реченице за српски језик. За случај анотације корпуса потребно је проучити постојеће методе анотације корпуса. Наредни корак је прикупљање, анализирање и складиштење анкета ФТН-а у бази података. На основу коментара из анкета, потребно је формирати и анотирати корпус који ће бити коришћен за развој и евалуацију модела (такозвани корпус златног стандарда). Затим у наредном кораку је потребно развити неопходне алате и ресурсе за детекцију сегмената реченице (фраза и клауза) за српски језик. Након тога је потребно формирати модел за аутоматску анализу мишљења студената заснованог на машинском учењу и лексичким правилима.

Крајњи корак је евалуација развијеног модела помоћу анотираног корпуса анкета и дискусија добијених резултата.

Очекивани резултат истраживања докторске дисертације јесте емпиријски доказ полазне хипотезе и циљева који су из њих изведени, односно:

- формирање анотираног корпуса текстуалних коментара студентских анкета намењеног за развој и евалуацију модела за аутоматско издвајање мишљења,
- развијање алата и ресурса за детекцију сегмената реченице (фраза и клауза) за српски језик,
- формирање модела за аутоматску анализу мишљења студената заснованог на машинском учењу и лексичким правилима.
- формирање модела за аутоматску анализу мишљења студената заснованог на дубоком учењу.

Очекивани модел има широке могућности примене у свим високошколским установама где је неопходно формирање извештаја и класификација података према сентименту и аспекту. Конкретни примери где би овакав систем био од посебне користи су наведени у наставку.

Први пример конкретне примене су сумарни извештаји базирани на резултатима аспектне и сентимент анализе. У зависности од типа, анкете углавном садрже опште податке о анкетираној особи попут смера студирања, године уписа, године студирања, анкетираном предмету итд. На основу анализе сентимента и аспеката у коментарима анкета, могу се формирати извештаји који наводе следеће:

- да ли су анкетирани студенти опште задовољни или нису задовољни,
- којим аспектима студирања су студенти највише задовољни, односно нису задовољни,
- којим аспектима студирања су студенти највише задовољни, односно нису задовољни, на нивоу одређеног смера, предмета, године студирања, године уписа, или наставника,
- промена задовољства, односно незадовољства, током анкетираних година за одређене аспекте студирања за одређени смер, предмет, наставника или годину студија.

Други пример конкретне примене резултата аспектне и сентимент анализе са техникама анализе података. На пример, помоћу техника кластеровања се могу утврдити да ли постоје групе сличних смерова или предмета или наставника по одређеном сентименту или аспекту студирања. На основу тога управа високошколске установе може уочити који су то кључни аспекти чиме су студенти задовољни, односно

нису задовољни, и спровести додатне акције у циљу побољшања квалитета студирања. На пример, може утврдити постојање корелација између сентимента и оцена студената на одређеном предмету, током одређеног периода. Тиме се може предвидети оцена студената за наредни период, и у случају негативног тренда, обратити већа пажња ка оним аспектима којим студенти нису задовољни. Са друге стране, може се обратити већа пажња на аспекте чиме су студенти задовољни и онда приказати какав профил смера или предмета или организације доводи до великог задовољства и успеха студената.

1.3 Структура дисертације

Ова дисертација је организована у осам поглавља. У првом поглављу је описан предмет истраживања и мотивација за истим. Дефинисани су циљеви истраживања и хипотезе које су представљале оквир ове докторске дисертације. Након уводног поглавља (прво поглавље), су описане теоријске основе области докторске дисертације. Дат је опис основних појмова обраде природног језика, сентимент анализе и аспектно базиране сентимент анализе. У наставку су описан општи методолошки оквир за развој система за природну обраду језика. Затим су описани приступи анализе мишљења у погледу различитих типова модела. Прво су описани иницијални приступи обраде природног језика коришћењем модела заснованих на речницима и правилима. Наредни корак у приступима представљају модели машинског учења који су груписани у две групе – групу класичних модела машинског учења и групу модела дубоког учења. Модели дубоког учења представљају најактуелније приступе у области обраде природног језика, те су стога описани детаљи најпознатијих модела и начина обучавања и оптимизације модела. Велики акценат је стављен и на обучене језичке моделе (енгл. *pre-trained language models*) који су допринели великом напретку у истраживању у области обраде природног језика.

У трећем поглављу су описани извори релевантних за обраду природног језика и анализу мишљења, са фокусом на аспектно базирану сентимент анализу. Велики акценат је стављен на типове коришћених модела, ниво анализе (документ, реченица, сегмент реченице) и постигнуте резултате евалуације модела. Прво су описани првобитни приступи засновани на језичким ресурсима (речници, правила, онтологије) а затим приступи засновани на машинском учењу.

У четвртом поглављу је описан процес формирања корпуса златног стандарда који представља један од резултата ове докторске дисертације. Описан је процес анотације корпуса атрибутима потребним за извођење аспектне сентимент анализе. У

наставку је извршена анализа поузданости анотатора и описани су статистички детаљи корпуса. Упоредно је описан и други корпус коментара високог образовања који је резултат сродног истраживања а користио за упоредну евалуацију модела развијеног о овој дисертацији.

У петом поглављу је описана методологија имплементације система за аутоматску анализу мишљења у коментарима студената високог образовања. Систем за аспектно базирану сентимент анализу је заснован на засебним модулима за идентификацију аспеката студирања и сентимент поларитета, чији резултати се интегришу у један коначни резултат. Експериментисано је са више стратегија интеграције резултата модела у оквиру модула и одабрани су крајњи модели на основу перформанси на валидационом скупу.

У шестом поглављу су описани експериментални резултати развијеног модела за аспектно базирану сентимент анализу. Извршена је анализа грешака и дискусија резултата крајњих модела. Дускутоване су и разлике између два корпуса и две групе модела како би се образложили добијени резултати модела. Такође је дискутована примењивост предложеног модела на друге високошколске установе у Србији и описана су ограничења овог истраживања.

У седмом поглављу је описана примена резултата аспектно базиране сентимент анализе. Представљени су типови извештаја који руководству високошколских установа могу бити помоћ при доношењу наредних пословних одлука у циљу повећања квалитета студирања и студијских програма. Поред генералних извештаја на нивоу факултета, описани су специфични извештаји који могу дати бољи увид у задовољство студената одређеним аспектом студирања. Увођењем временске линије у извештаје се може испратити промена сентимента у односу на одређени аспект.

У осмом поглављу је наведен закључак ове дисертације која поред осврта на кључне детаље овог истраживања даје и наредне кораке за будући рад из области аспектно базиране сентимент анализе на српском језику.

Након Закључка су наведени извори истраживања који су коришћени у овој докторској дисертацији, прилози и биографија аутора ове докторске дисертације.

2 Теоријске основе

Предмет ове докторске дисертације је обрада природног језика са задатаком аспектно базиране сентимент анализе који припада широј области анализе мишљења тј. утврђивању сентимента у природном језику. Циљ аспектно базиране сентимент анализе је утврђивање аспеката као носиоца исказаног сентимента. Стога ће прво бити изложени основни појмови обраде природног језика, затим анализе мишљења са фокусом утврђивања аспеката и сентимент поларитета. Након тога је приказана генерална методологија за решавање задатака обраде природног језика која укључује низ стандардних корака, од постављања проблема, дефинисања корпуса и обучавања и евалуације модела.

2.1 Аутоматска обрада природног језика

Природан језик по ужој дефиницији представља начин комуникације људи, најчешће усменим или писменим путем. Говорна комуникација се врши употребом природног језика, који врло често бива сведен на писани облик. Природни језик у писаном облику представља текст, скуп речи који су организовани по одређеној граматици и имају одређено значење. Под појмом обраде природног језика се подразумева аутоматска анализа природног језика уз помоћ машине тј. рачунара. Почетни кораци у *NLP* области су учињени пре више од пет деценија у научној области лингвистике, која се касније проширила и на друге научне области појављивањем персоналних рачунара.

Лингвистика је научна област истраживања људског језика која обухвата разне области попут морфологије, синтаксе, фонологије, семантике, лексилогије и тако даље. Особе које се баве овом науком се називају теоријским лингвистима. Појавом друге генерације рачунара је настала и међудисциплинарна област рачунарска лингвистика која се бави рачунским моделовањем природног језика у циљу прављења компијутерских програма за решавање различитих лингвистичких питања. На пример,

средином педесетих година прошлог века[3] је први пут јавно демонстрирано врло ограничено машинско превођење са руског на енглески језик (Hutchins, 1999).

Тежња рачунарских лингвиста је да се разумевање природног језика сведе на математичке формулације, фокусирајући се на тестирање граматике коју су истражили теоријски лингвисти. Ширењем ове области истраживања су настале две научне гране - теоријска и примењена рачунарска лингивистика. Теоријска рачунарска лингвистика се бави развојем формалних теорија граматике (парсирањем) и семантиком, док је фокус примењене рачунарске лингвистике на практичним исходима употребе моделовања природног језика. Термин рачунарска лингвистика се данас узима као синоним за *NLP*, који заједно припадају људској језичкој технологији (енгл. *Human Language Technology*).

Према дефиницији аутора (Goldberg, 2017), *NLP* подразумева развијање метода и алгоритама које за улаз или излаз стварају текстуалне податке које чине природни језик. Природан језик представља специфичан тип података чија обрада представља изазован задатак. Аутор (Goldberg, 2017) истиче да је природан језик врло двосмислен, варијабилан и променљив. Једна иста информација се може написати на више начина употребом различитих речи. Исто тако, променом једне речи или заменом места речи у реченици се добија други смисао и друго значење. Језик се попут људи који га користе стално мења и евалуира. Природни језик који је коришћен у прошлом веку се знатно разликује од данашњег у погледу нових речи, појмова, израза и граматичких структура.

Обрада природног језика је до деведесетих година прошлог века била базирана на класичним приступима заснованих на правилима, који су замењени статисткичким приступима заснованим на машинском учењу. Доступност велике количине текстуалних података и рачунарских система велике процесорске моћи је омогућило да се нове и другачије ствари могу истражити веома брзо писањем и покретањем софтвера. Данас, статистички приступи обради природног језика су највише базирани на коришћењу дубоких вештачких неуронских мрежа због постизања одличних резултата на већини *NLP* задатака и због развоја робусних система за специфичну намену (нпр. машинско превођење, сумаризација природног језика, генерисање природног језика, разумевање природног језика, задатак одговора на питање).

У оквиру *NLP* задатака специфичне намене се користе и други задаци који се појављују у истраживању *NLP* области. Иако се *NLP* задаци међусобно преплићу они се могу груписати на основу области истраживања лингвистике попут морфологије, синтаксе, фонологије, семантике, и тако даље. Задаци који се истражују у групи морфолошке анализе се односе на сегментацију текста на морфеме (најмања јединица

[3] *Georgetown–IBM* експеримент: https://www.ibm.com/ibm/history/exhibits/701/701_translator.html

која има смисао), свођење речи на основни облик - стемовање (енгл. *stemming*) и лематизација (енгл. *lemmatization*), и означавања типа речи (енгл. *Part-Of-Speech - POS*). У групу синктаксичке анализе се убрајају задаци препознавања граница реченица и синктасичког парсирања, односно одређивања граматичке структуре реченица. Задаци препознавања и сегментације говора, генерисања текста на основу говора припадају групи фонологије. Задаци попут препознавања именованих ентитета (енгл. *Named Entity Recognition*) и сентимент анализе припадају групи лексичке семантике, која се односи на анализу индивидуалних речи. Групу семантике релација чине задаци семантичког парсирања, односно одређивања семантичких релација речи у реченицама.

2.2 Анализа мишљења – сентимент анализа

Исказивање мишљења односно става у вези одређене ствари је повезано са исказивањем субјективног осећања и уверења (Liu, 2015). Према психологији и речима аутора (Liu, 2015), осећања имају централну улогу у људском понашању. Људски избори и одлуке које доносе су у великој мери зависне од њиховог осећања и расположења у датом тренутку. Док су уверења људи изграђена на основу окружења односно уверења других људи и њиховог погледа на свет. Под окружењем људи се поред физичког подразумева и виртуелно окружење где људи комуницирају. Стога се може закључити да људи по природи истражују мишљења и осећања других људи везано за одређену тему која их интересује.

Анализа мишљења и сентимент анализа се у литератури појављују као синоними који за циљ имају издвајање мишљења и сентимента из текста природног језика коришћењем рачунарских метода. Под термином мишљења се сматрају било какве изјаве и ставови у вези неког објекта. Термин сентимента се односи на позитивно или негативно осећање исказано мишљењем. Често се у сентимент убраја и неутрално осећање које означава одсуство и позитивног и негативног осећаја.

Појавом већег броја интернет страница и платформи (нпр. форуми, друштвене мреже) се повећао број текстуалних података о исказаном мишљењу у дигиталном облику. Људи су почели да се повезују једни са другима и да на слободан начин износе своја мишљења и ставове о било којој теми без обзира на њихову географску локацију. Тиме се од 2002. године повећао број истраживања у области сентимент анализе у друштвеним мрежама са циљем издвајања корисних информација о чему људи причају и која су њихова мишљења (Liu, 2015). Последњих година је примена анализе мишљења и сентимента проширена на скоро све домене, од потрошачких производа, мултимедије, финансија, социјалних догађаја, политике итд. Добијене информације се,

на пример, могу искористити за апликације попут система за преоруку производа и сервиса или утврђивање која политичка партија или кандидат ће победити на изборима.

Према аутору (Liu, 2015), мишљења људи су веома битна за појединце, предузећа и организације. Предузећа и организације желе да сазнају ставове корисника и јавности о њиховим производима и услугама. На основу ових мишљења се могу донети ефикасније одлуке у вези одржавања или побољшања одређених услуга или производа. Појединици, као потрошачи, такође желе да сазнају мишљења других у вези одређених производа и услуга пре него што се одлуче да купе дати производ или услуге. Велика пажња анализе мишљења је усмерена и на политичке партије у периоду избора када је потребно испитати ставове и расположење јавног мњења у вези одређених питања и догађаја.

Пре настанка поменутих интернет страница се у прошлости мишљење прикупљало усменим путем. Појединци су за мишљење других питали своју породицу и пријатеље. Предузећа и организације су мишљења својих корисника прикупљали путем сопствених анкета чије анализирање је представљало дуг и мукотрпан посао, подложан грешкама људског фактора. Настанком интернет форума и друштвених мрежа се до тражених мишљења доста лакше долазило. Међутим, у случају да је потребно прибавити мишљења строго контролисане групе (нпр. по годинама, полу, образованом нивоу или послу) онда се и даље спроводе анкете и упитници јер идентитет на интернет форумима и платформама за друштвене мреже није засигурно тачан. Како би се сумирало свеобухватно мишљење и пронашло решење одређеног проблема, потребно је анализирати оба извора, јавни на интернету и приватни у оквиру предузећа или организације.

Исказивање мишљења и сентимента је субјективне природе, стога се реченице које садрже нечије мишљење или сентимент називају субјективним реченицама. Насупрот субјективним реченицама постоје и објективне реченице, које садрже чињенице. Иако објективне реченице не садрже субјекат, оне могу имати позитиван или негативан сентимент. На пример, реченица „Данас је сунчан дан.“ је објективна реченица која садржи чињеницу исказану у позитивном сентименту. Оваква реченица се може тумачити да је аутор исте имао позитивне осећаје приликом писања. Стога се анализа мишљења и сентимента може вршити и над субјективним и објективним реченицама (Liu, 2015).

Приступи анализе мишљења се заснивају на методама базираним на знању, статистичким методама и хибридним методама. Методе базиране на знању идентификацију сентимента врше на основу експлицитних речи и фраза које носе одређени сентимент. Сентимент речи и изрази се идентификују применом речника, правила, онтологија и других типова репрезентација знања. Софистицираније приступ

идентификације сентимента се заснива на анализи граматичких релација у тексту применом статистичких метода. Статистичке методе су базиране на машинском учењу попут модела Наивног Бајеса (енгл. *Naive Bayes - NB*), машине потпорних вектора (енгл. *Support Vector Machines - SVM*), дубоког учења (енгл. *deep learning*), и тако даље. Хибридне методе комбинују методе машинског учења и различитих репрезентација знања попут онтологија и семантичких мрежа, како би уочили сентимент у контексту који не садрже експлицитне сентимент изразе али су на неки начин повезани са другим контекстима која садрже.

Анализа мишљења и сентимента је према литератури највише вршена на три нивоа гранулираности: ниво документа, ниво реченице и ниво аспекта или дела реченице. Највиши ниво анализе је ниво документа где се анализира цео документ мишљења. Документ, на пример, може бити рецензија производа са интернет продавнице, где се одређује да ли рецензија изражава свеукупно позитивно или негативно мишљење о одређеном производу. У овом случају се претпоставља да цела рецензија, односно докумнет, изражава мишљење о једном ентитету (нпр. производ, услуга). Међутим овај ниво анализе није погодан за случај када се у документу помиње више ентитета. Следећи ниво анализе је ниво реченице, где се одређује да ли цела реченица изражава позитивно или негативно мишљење. Овај ниво анализе је финије гранулације тј. даје детаљније информације него анализе на нивоу документа. На овај начин се може извршити издвајање реченица према сентимент поларитету. Међутим, ни анализа мишљења на нивоу реченице није довољна да се установи о чему се ради у сваком мишљењу тј. којим објектом људи јесу или нису задовољни. На пример, за следећу реченицу „Настава се добро одржала али недостаје књига.“ нема неког смисла одредити сентимент на нивоу реченице јер се изражава позитивно мишљење о настави а негативно о метеријалима (нпр. књиге, скрипте, презентације). Да би се постигли ови прецизнији резултати потребно је анализирати мишљење на нивоу аспекта, који се у новијој литератури назива аспектно базирана сентимент анализа.

2.2.1 Пред-процесирање текста

Неизоставни кораци пре примене одређеног *NLP* модела су обрада података, издвајање језичких особина (карактеристика) и формирање репрезентације текста. У овом делу ће бити описан начин на који се може од сировог текста доћи до оног облика који се користи за формирање *NLP* модела.

2.2.1.1 Грубо пред-процесирање

Текстуални подаци се прикупљају из разних извора попут интернет страница, дигиталних докумената и доменских система. Потребно је уклонити све сувишне податаке за *NLP* као што су интернет тагови, линкови, заглавља докумената, специјални знакови, емотикони, дијакритичке знакове, величине слова и тако даље. Уколико су емотикони у подацима заступљени у довољној мери онда се могу изузети из процеса уклањања у сврху анализе мишљења.

Полазни корак пред-процесирања дужих текстова попут докумената и пасуса подразумева технику сегментације реченица, односно дељење докумената и пасуса на реченице. Најједноставнија сегментација реченица се заснива на знаковима интерпункције попут тачке, упитника, узвичника, и тако даље. По потреби се се сегментација може наставити на дељење реченица на сегменте реченице попут фраза и калуза. Клаузе садрже субјекат и предикат, док су фразе мање целине од клауза и не садрже везу између субјекта и предиката. У овом случају је поред знака интерпункције (нпр. зареза) потребно узети у обзир резултате наредног корака (Секција 2.2.1.2) а то су језичке особине.

Наредна техника пред-процесирања текста је токенизација (енгл. *tokenization*) и подразумева дељење текста на речи односно токене. Токенизација се уобичајено извршава по размацима између речи где је потребно узети у обзир знакове интерпункције. Скуп знакова интерпункције и слова у одређеним анализама текста имају значење попут емотикона (нпр. :-), :-D), хеш тагова (нпр. *#happy, #perfect*), и тако даље. Резултат токенизације текста је низ токена, односно речи. У наставку се примењује техника свођења речи на основни облик тј. стемовање (енгл. *stemming*). Техника стемовања је најједноставнији облик свођења речи на основни облик (енгл. *stem*) где се уклањају завршеци, односно суфикси, речи. У овој техници сведени облик речи може бити неисправан морфолошки корен речи, али може бити довољан за свођење сродних речи на исти основни облик. Комплекснија техника свођења речи на основни облик је лематизација (енгл. *lemmatization*) која узима у обзир контекст и примењује морфолошку анализу речи. У овом случају, леме тј. коренски облик речи је морфолошки тачан. За ову технику су потребне информације о типу речи који се добијају у наредном кораку финијег пред-процесирања (Секција 2.2.1.2).

Веома чест корак пред-процесирања текста за случај анализе мишљења је уклањање најфреквентнијих речи у тексту које не носе сентимент, познате и под термином стоп речи (енгл. *stop words*) тј. врсте речи попут предлога, везника и речца. Пример наведених корака је дат у наставку за коментар на енглеском и српском језику. Коментар на енглеском језику „*Everyone is talking about artificial intelligence these days!*

😊 *#ai #hype* “, се пред-процесирањем са кораком лематизације своди на низ токена спојених размаком „*everyone be talk artificial intelligence these day* “. Исти пример коментара на српском језику „Данас сви причају о вештачкој интелигенцији! 😊 *#ai #hype* “, се своди на низ токена спојених размаком „данас сав причати о вештачки интелигенција “. У зависности од домена проблема, након овог корака се може приступити финијем пред-процесирању података пре формирања репрезентације текста (Секција 2.2.1.3) која одговара формирању *NLP* модела.

2.2.1.2 Фино пред-процесирање

У кораку финог пред-процесирања података се приступа техникама за издвајање језичких особина попут типова речи, веза између типова речи, граматичких структура, и тако даље. У наставку су описане технике граматичког означавања и парсирања текста.

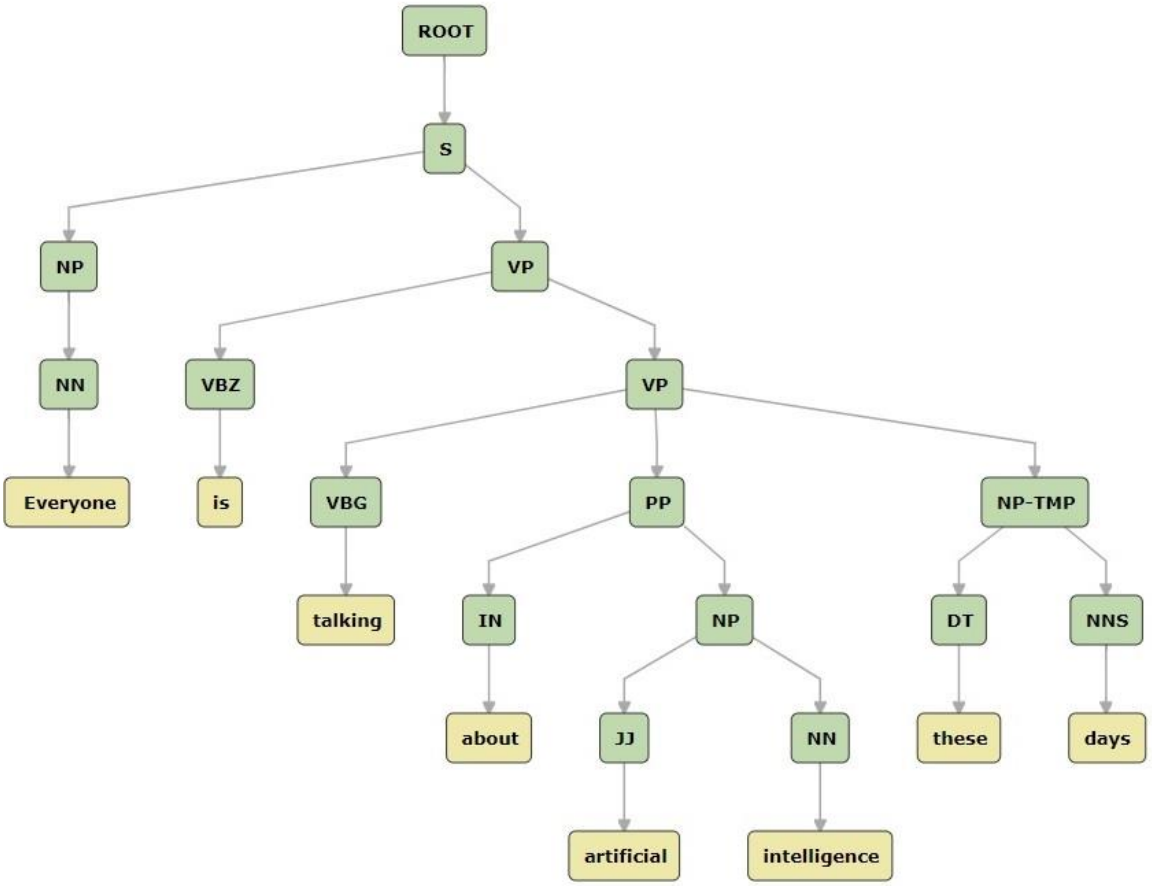
Техника граматичког означавање речи у реченици се односи на процес додељивања врсте речи (енгл. *part-of-speech - POS*) у зависности на њихову дефиницију и контекст. Свака реч у реченици у српском језику се тагује (енгл. *tagging*) тј. означава једном врстом речи попут именице, придева, глагола, прилога, предлога, заменице, везника, узвика, речце и броја. За морфолошки богате језике (нпр. српски језик) се поред врсте речи граматичким означавањем придодају и додатни подаци попут типа, пола, броја и придева. Пример граматичког означавања речи у реченици на енглеском и српском језику, након корака лематизације, је дат у наставку. Речи из реченице на енглеском језику „*Everyone is talking about artificial intelligence these days* “, се своде на основни облик и тагују скраћеницама одговарајућих врста речи „*everyone/NN be/VB talk/VB about/IN artificial/JJ intelligence/NN these/DT day/NN* “. Исти пример реченице на српском језику „Данас сви причају о вештачкој интелигенцији“, речи се своде на основни облик и тагују скраћеницама одговарајућих врста речи „данас*/RB* сав*/JJ* причати*/VB* о*/IN* вештачки*/JJ* интелигенција*/NN* “. Скраћенице врста речи у датим примерима се односе на именице (енгл. *noun - NN*), глаголе (енгл. *verb - VB*), придеве (енгл. *adjective - JJ*), предлоге (енгл. *preposition - IN*), прилози (енгл. *adverb - RB*) и одреднице (енгл. *determiner - DT*). Применом граматичког означавања без претходног свођења речи на основни облик се добијају детаљније информације о врстама речи. На пример у српском језику, за именице детаљније информације су о типу, роду, броју, падежу. За придеве се добијају детаљније информације о типу, глаголској форми, времену, лицу, роду, броју, стању, и тако даље.

На основу врсте речи се може претпоставити тј. одредити вероватноћа суседних речи и синтаксичка структура фраза (Jurafsky and Martin, 2009). На пример, ако је врста

одређене речи именица, знамо да су именице део именичких фраза и да за речи пре именице можемо очекивати врсте речи попут придева и глагола. Граматичко означавање, познатије и као *POS* таговање, чини кључни део синтаксичког парсирања реченица и одређивања именованих ентитета (енгл. *named entities*), које је кључно за *NLP* задатке попут издвајања информација.

Техника парсирања (енгл. *parsing*) у *NLP* области подразумева процес утврђивања синтаксичке структуре текста анализирањем његових саставних речи на основу граматике за одређени језик. Синтаксичка структура текста подразумева тип, облик, улогу и међусобну повезаност речи у тексту. У литератури се парсирање јавља и под појмом синтаксичке анализе (енгл. *syntactic analysis*) и анализе синтаксе (енгл. *syntax analysis*). Једноставнији приступи парсирања се базирају на примени основних образаца претраживања (нпр. попут регуларних израза). Док су комплекснији приступи узимају у обзир контекст реченице чиме се задржавају семантичке везе између делова реченице (Jurafsky and Martin, 2009).

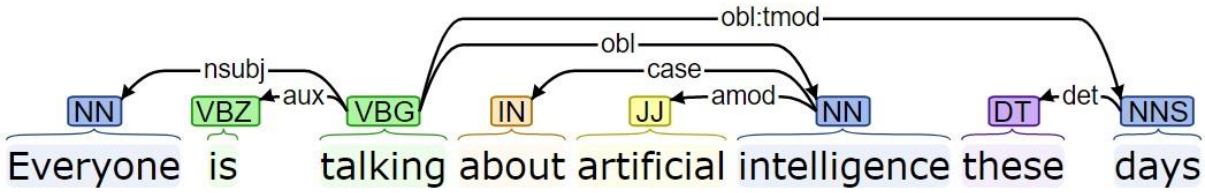
Према граматици која се користи при парсирању, односно синтаксичкој структури која се добија парсирањем, постоје две врсте парсирања. Парсирање засновано на безконтексној граматици (егнл. *context-free grammars*) се фокусира на идентификацију и класификацију фраза и њихове структуре. Један тип ове врсте парсирања је плитко парсирање, познатије по жаргонском изразу „комадања“ (енгл. *chunking*) реченице на комаде тј. делове реченице. Плитко парсирање реченице је процес у ком се *POS* таговањем добију врсте речи који се затим повезују на вишем граматичком нивоу попут именичких, глаголских, предлошких фраза, формирајући тиме делимично стабло. Дубља хијерархијска структура тј. потпуно стабло парсирања се добија потпуним или дубоким парсирањем (енгл. *deep parsing*). Код структуре потпуног стабла парсирања корен стабла је реченица, средњи чворови стабла су врсте речи попут иманица, глагола, придева, итд., а крајњи чворови (листови) су речи из коренске реченице. Графички приказ потпуне структуре парсирања реченице на енглеском језику је дата на Слика 1.



Слика 1 *Потпуно стабло парсирања засновано на безконтекстној граматици*[4]

Друга врста парсирања је заснована на граматици зависности (енгл. *dependency grammar*) где је фокус на идентификацији зависности речи. Структура фраза у овој врсти парсирања не игра велику улогу већ се синтаксичка структура реченице описује речима и одговарајућим скупом граматичких релација између речи (Jurafsky and Martin, 2009). Структура зависности се добија потпуним тј. дубоким парсирањем док случај плитког парсирања није чест. Графички приказ структуре зависности добијене парсирањем реченице на енглеском језику је дата на Слика 2.

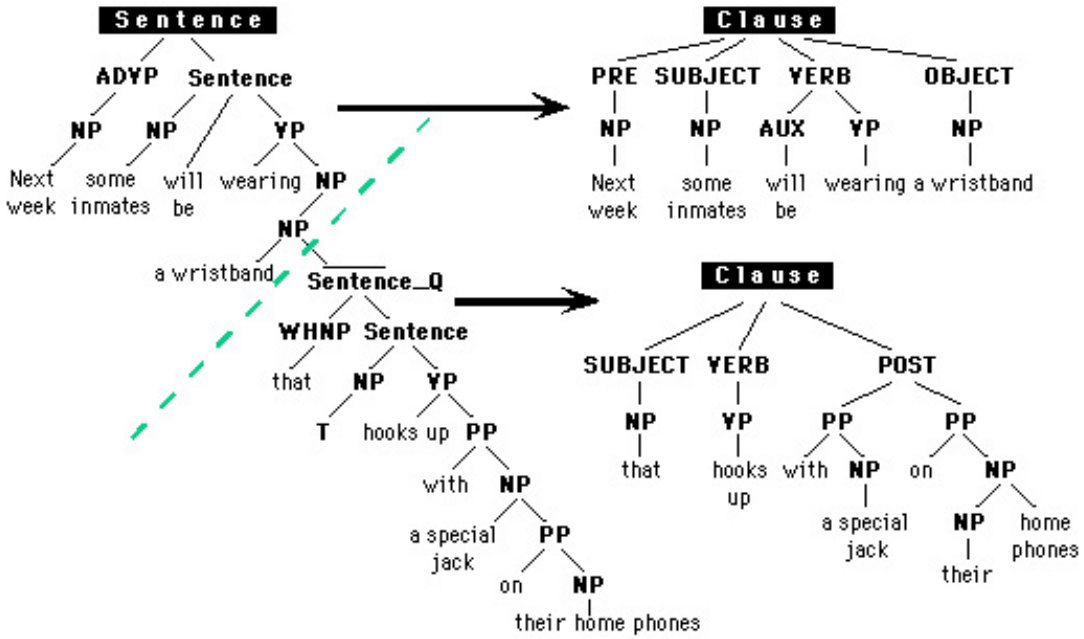
[4] Графички приказ стабла добијен *Stanford CoreNLP* алатом доступном на адреси: https://corenlp.run/



Слика 2 *Стабло зависности парсирања засновано на граматици зависности*[4]

Разлике између структуре фраза и структуре зависности се огледа у броју чворова који одговарају речима из реченице. Код структуре фраза за сваку реч постоји један или више чворова док се код структуре зависности за сваку реч издваја тачно један чвор стабла.

Пример сегментације реченице на клаузе (Секција 2.2.1.1) базирано на информацијама структуре реченице је дат на Слика 3. За пример сложене реченице на енглеском језику *„Next week some inmates will be wearing a wristband that hooks up with a special jack on their home phones.“* се формира потпуна структура реченице. На основу граматичких правила се дефинише позиција дељења реченице на клаузе и фразе. Тачније, на основу потпуне структуре парсирања се формирају две повезане под структуре као на Слика 3.



Слика 3 *Сегментација реченице базирана на дубоком стаблу парсирања*[5]

[5] Слика преузета са адресе: http://alumni.media.mit.edu/~cahn/loq/loq--ling.html

2.2.1.3 Репрезентација текста

У овој секцији су описани начини репрезентације текста од којих многи користе резултате грубог (Секција 2.2.1.1) и финог пред-процесирања података (Секција 2.2.1.2). Утицај пред-процесирања на репрезентацију текста се своди на канонизацију текста (енгл. *canonicalization*), односно свођење текста на што општију каноничку форму чиме се смањује фреквенција термина у целом скупу података. Формирање репрезентације текста је један од најважнијих корака у формирању *NLP* модела.

Текстуална секвенца се може репрезентовати као низ карактера, низ речи, низ нумеричких вредности у векторском простору и низ нумеричких вредности са проширеном семантиком у векторском простору. Тип репрезентације текста зависи од тога да ли текстуалну секвенцу чини једна реч или више од једне речи (клауза, фраза, реченица, документ).

2.2.1.3.1 Репрезентација речи

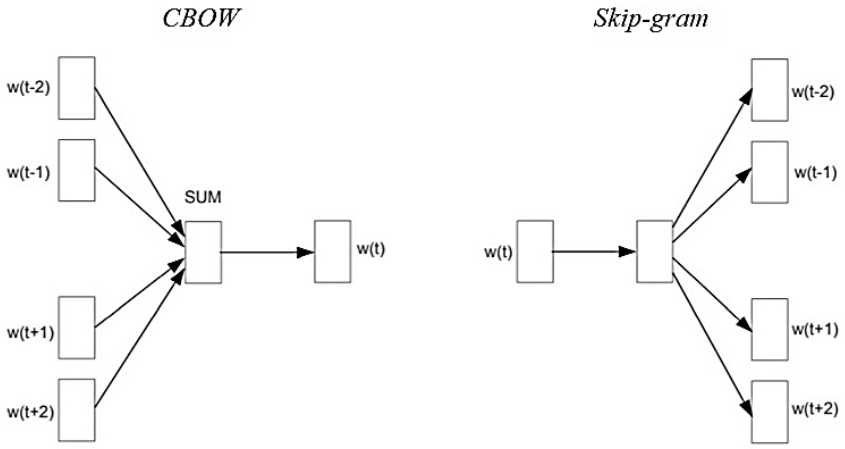
Реч се у традиционалним *NLP* приступима најчешће репрезентује као вектор јединице (енгл. *one-hot-vector*), односно вектором димензије речника где се на позицији посматране речи налази број један. Речник чини скуп одабраних речи из одређеног корпуса. Мана репрезентације речи вектором јединице је немогућност утврђивања симетричности са векторима јединица других сличних речи. На пример, речи „наставник“ и „професор“ су веома сличног значења која се не може утврдити на основу репрезентације речи вектором јединице. Делимично решење утврђивања симетричности се може пронаћи у лексичким базама речи и семантичких релација, попут *WordNet*[6]-а. Мане ових речника су непотпуност појмова, субјективност додељених семантичких релација, време формирања и прилагођавања зависи од људског фактора.

Веће значење речи дају речи које се често појављују у њеној близини тј. контексту. Стога се у нешто другачијем приступу речи тј. термини репрезентују као вектор нумеричких вредности у векторском простору (енгл. *word embeddings*) чије вредности зависе од речи које се често појављују у истом контексту. Један од првих

[6] Енциклопедија језика и лингвистике за енглески језик: https://wordnet.princeton.edu/

приступа векторске репрезентације речи је *Word2Vec* (Mikolov *et al.*, 2013) модел који користи неуронску мрежу (Секција 2.5.3.2.1) за учење веза између речи из великог корпуса текста. Свака реч у речнику је репрезентована вектором нумеричких вредности фиксне дужине. Идеја се базира на рачунању сличности вектора посматране речи и вектора речи контекста тј. одређен број речи пре и после посматране речи. Резултат мере сличности вектора су нумеричке вредности које представљају вероватноће да се посматрана реч нађе у контексту осталих речи из речника. Неуронска мрежа *Word2Vec* модела се обучава са циљем максимизовања ових вероватноћа.

Постоје две варијанте *Word2Vec* типа модела које се разликују по броју речи које се предвиђају (Слика 4). *Skip-gram* модел за посматрану реч предвиђа контекст тј. речи са највећом вероватноћом. Обрнут поступак предвиђања врши *CBOW* (енгл. *Continuous Bag of Words - CBOW*) модел. За посматран контекст се предвиђа централна реч која има највећу вероватноћу. *Skip-gram* модел се показао ефикаснијим за ретке речи у корпусу док се *CBOW* модел показао бољим за фреквентне речи у корпусу. Међутим, постоји неколико мана *Word2Vec* типа модела. Комплексност модела и дужина обучавања модела се скалира према величини корпуса. Са мањим корпусом се добијају лошије перформансе предвиђања речи и контекста. Такође, не чувају се глобалне статистичке информације корпуса него само семантичке реч-контекст аналогије.



Слика 4 *Word2Vec модел*

За разлику од *Word2Vec* модела, *GloVe* (Pennington *et al.*, 2014) модел поред аналогије вектора речи чува и глобалне статистичке информације корпуса. Идеја се базира на формирању матрице појављивања термина за сваки термин у целом корпусу (енгл. *co-occurrence matrix*). Матрица се формира за околину тј. контекст посматране речи. Пример матрице појављивања термина за корпус који се састоји од две реченице „Ja volim NLP.“ и „Ja volim duboko učenje.“, је приказана у Табела I.

Табела I *Матрица појављивања термина у корпусу*

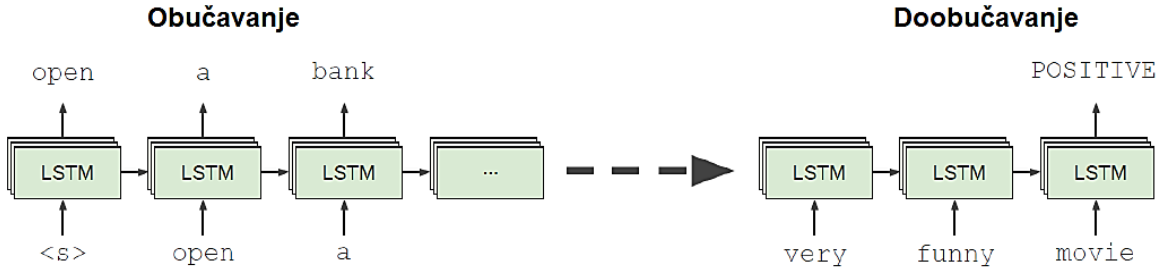
	JA	**VOLIM**	**DUBOKO**	**UČENJE**	**NLP**	**.**
JA	0	2	0	0	0	0
VOLIM	2	0	1	0	1	0
DUBOKO	0	1	0	1	0	0
UČENJE	0	0	1	0	0	1
NLP	0	1	0	0	0	1
.	0	0	0	1	1	0

С обзиром на то да је резултујућа матрица појављивања термина у великом корпусу великих димензија, приступа се смањењу димензија применом декомпозиције јединствених вредности (енгл. *Singular Value Decomposition - SVD*). Матрица се декомпонује на каноничку форму сингуларних вредности (корена) и сингуларних вектора. Задржавањем већих а анулирањем мањих сингуларних вредности се постиже апроксимација полазне матрице тј. смањење димензија матрице. Обучавање *GloVe* модела је доста брже и остварене су добре перформансе и са мањом величином корпуса. Међутим, и *GloVe* и *Word2Vec* модели не разматрају морфолошку структуру речи и ограничени су на речи које се појављују у корпусу током обучавања.

Захваљујући морфолошком приступу репрезентације речи *fastText* (Bojanowski *et al.*, 2017) модела је омогућено формирање векторске репрезентације речи које се нису појавиле током обучавања. Овај модел је базиран на *Skip-gram* моделу где је свака реч репрезентована сумом векторских репрезентација њених подречи тј. делова сачињених од низа карактера одређене дужине. На пример, енглеска реч „*where*“ се за дужину подречи три посматра као низ „*wh, whe, her, ere, re*“, чије векторске репрезентације се сумирају и представљају векторску репрезентацију речи „*where*“. Међутим, *Word2Vec*, *GloVe* и *fastText* модели не узимају у разматрање редослед речи контекста што доводи до губитка синтаксе и семантике реченице у којој се посматра одређени контекст. То значи да би векторска репрезентација речи била иста за реченице са различитим контекстом. На пример, векторска репрезентација речи „град“ би била иста за реченице „Данас је падао град.“ и „Нови Сад је леп град.“.

ULMFiT (Howard and Ruder, 2018), *ELMo* (Peters *et al.*, 2018) и *BERT* (Devlin *et al.*, 2018) модели користе дубоке неуронске мреже (Секција 2.5.3.2.3) различитих архитектура за учење векторских репрезентација речи. Ови модели узимају у разматрање редослед речи контекста што доводи до тога да за једну исту реч постоје више различитих векторских репрезентација у зависности од контекста. Језички модели *ULMFiT* и *ELMo* су формирани од *LSTM* јединица рекурентне мреже (Секција 2.5.3.2.3.2) и обучавани на великим неанотираним корпусима са задатком предвиђања наредне речи у секвенци.

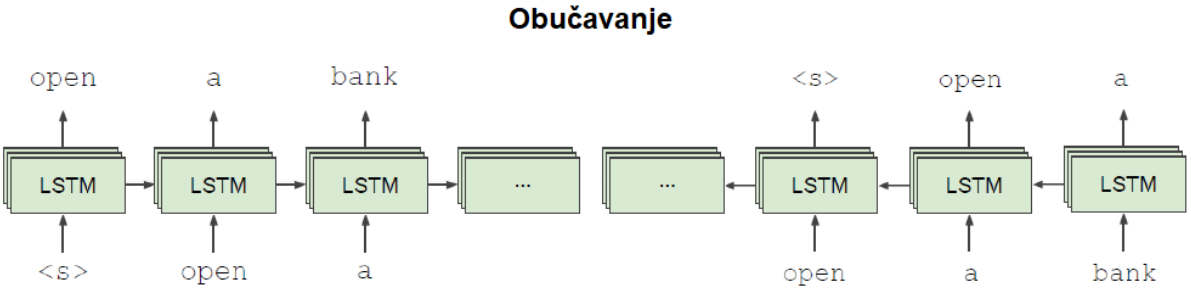
ULMFiT модел је формиран од трослојне *LSTM* мреже која је обучена на неанотираним подацима из корпуса генералног домена. На улаз модела су постављане речи из секвенце а излази су евалуирани над истом секвенцом речи померених у лево. Обучавањем на овај начин на великим корпусима се постиже веома добра језичка репрезентација која је потребна као основа за све задатке у *NLP* области. Аутори *ULMFiT* модела су закључили да се овакав језички модел може успешно дообучити на корпусу одређеног домена за конкретан *NLP* задатак, на пример сентимент анализе. Пример обучавања и дообучавања *ULMFiT* модела је приказан на Слика 5.



Слика 5 *Пример обучавња и дообучавања ULMFiT модела*

Поступак преузимања тежина језичког модела обученог на одређеном домену и дообучавање на другом домену или задатаку се назива трансфер знања (енгл. *transfer learning*). Дообучавање језичког модела се заснива на обучавању последњих слојева неуронске мреже док је већи део слојева замрзнут тј. тежине остају непромењене. Поступак трансфера знања је ублажио недостатак анотираних обучавајућих података тиме што се знање стечено за извршавање одређеног задатка може применити у решавању неког сличног задатка. Време формирања језичког модела је знатно скраћено, где је време потребно за обучавање на великом корпусу замењено знатно краћим временом за дообучавање модела за конкретан домен или задатак. Стога, поступак трансфера знања у *NLP* области је од 2017. године постао веома популаран.

Мана *ULMFiT* модела јесте да модел даје већу пажњу речима на почетку секвенце што у дужим секвенцама доводи до слабијег утицаја већег контекста на векторксу репрезентацију речи. У веома кратком временском периоду је предложен *ELMo* модел који поред *LSTM* мреже, чије су *LSTM* јединице повезане у смеру са лева на десно, додаје *LSTM* мрежу у супротном смеру (са десна на лево). На овај начин *ELMo* модел врши обучавање са две спојене секвенце у исто време. Пример обучавања *ELMo* модела је дат на Слика 6.



Слика 6 *Пример обучавња ELMo модела*

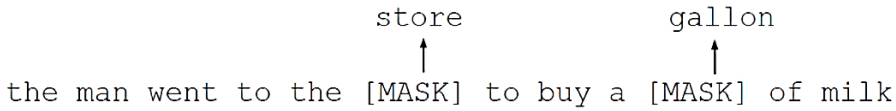
Попут *fastText* модела, *ELMo* модел посматра речи на нивоу карактера чиме се омогућава векторска репрезентација речи ван домена корпуса обучавања. Међутим, мана *ELMo* модела јесте што се иста пажња обраћа на све речи што у дужим секвенцама лоше утиче на квалитет векторске репрезентације речи. Захваљујући слоју пажње (Секција 2.5.3.2.3.4) и трансформер архитектури (Секција 2.5.3.2.3.5) дубоке вештачке неуронске мреже, представљен је *BERT* модел којим су постигнути до тада најбољи резултати у различитим *NLP* задацима (Devlin *et al.*, 2018).

За разлику од *ELMo* модела који садржи две засебне вишеслојне *LSTM* мреже за оба смера, *BERT* модел се базира на трансформер архитектури, двосмерној дубокој неуронској мрежи. По иницијалној форми трансформер архитектуре су укључени механизми енкодера и декодера. Међутим, циљ *BERT* модела је формирање језичког модела те је стога неопходан само механизам енкодера.

Обучавање *BERT* модела се врши на великим неанотираним корпусима са задатком предвиђања наредне речи у секвенци. Проблем двосмерног обучавање је што се за векторску репретентацију одређене речи користе информације о истој речи из другог смера. Из тог разлога *BERT* модел користи другачији приступ обучавања.

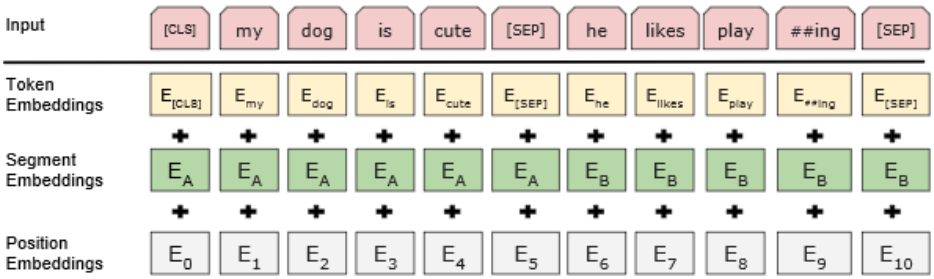
За обучавање *BERT* модела се користе две стратегије у циљу постизања бољих перформанси предвиђања. Циљ обучавања је минимизација комбиноване функције грешке за обе стратегије. Прву стратегију чини језички модел маскирања улазне секвенце (енгл. *Masked Langugage Model - MLM*). Одређени проценат насумично одабраних речи улазне секвенце механизма енкодера се маскира (замени) токеном маске (нпр. *[MASK]*). Затим се модел обучава да предвиди оргиналне вредности маскираних речи, анализирајући контекст тј. друге немаскиране речи у секвенци. При томе се води рачуна да проценат насумично одабраних речи не буде ни пуно велики ни пуно мали. Превише мали проценат маскираних речи представља комплексан циљ обучавања модела. Док превише велики проценат маскираних речи утиче на лошију репрезентацију контекста. Петнаест процената маскираних речи је, по ауторима *BERT* модела, оптималан однос између компексности циља обучавања и квалитета

репрезентације контекста. Задатак *MLM* модела на примеру једне секвенце на енглеском језику је приказан на Слика 7.



Слика 7 *Задатак маскирања улазне секвенце BERT модела*

У другој стратегији се обучавао модел који врши предвиђање наредне секвенце на основу дате улазне секвенце (енгл. *Next Sentence Prediction*). У овој стратегији су улаз трансформер модела чинили парови реченица на основу којих је предвиђано да ли друга реченица улазног пара реченица чини наредну реченицу прве реченице улазног пара реченица. У улазну секвенцу су додати токени почетка прве реченице (нпр. *[CLS]*) и токен раздвајања две реченице (нпр. *[SEP]*). Затим се улазна секвенца трансформише у секвенцу векторских репрезентација користећи сумарне информације о векторској репрезентацији речи, позицији речи у секванци и ознаке припадности првој или другој реченици. Пример репрезентације улаза *BERT* модела за задатак предвиђања наредне секвенце аутора (Devlin *et al.*, 2018) је дат на Слика 8.

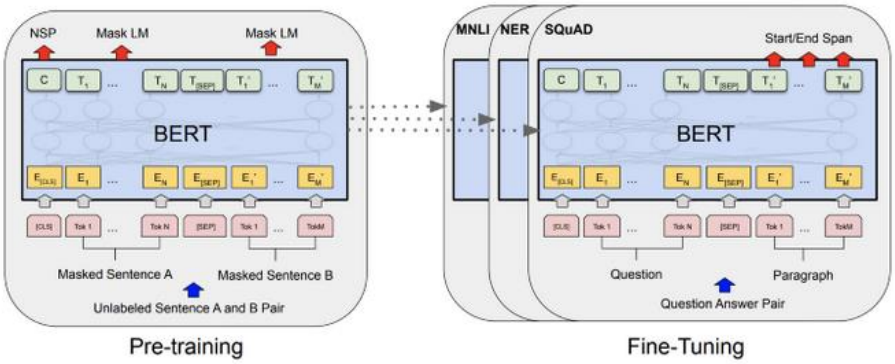


Слика 8 *Репрезентација улаза BERT модела*

За сваки токен улазне секвенце (укључујући и *[CLS]* и *[SEP]* токене) се преузима векторска репрезентација речи (нпр. E_{my}) користећи речник уобичајених речи и подречи (Wu *et al.*, 2016b). Подречи представљају делове речи, односно морфеме, као најмање семантичке јединице. На пример, енглеска реч „*unfortunately*“ се посматра као низ под речи „*un, fortunate, ly*“, чије засебне векторске репрезентације представљају векторску репрезентацију речи „*unfortunately*“. Стога реч „*playing*“ из примера Слика 8 се репрезентује са две засебне векторске репрезентације E_{play} и $E_{\#\#ing}$ две подречи „*play*“ и „*##ing*“. Затим се сваком токену улазне секвенце додељује векторска репрезентација ознаке припадности првој E_A или E_B другој реченици у улазној секвенци. У крајњем кораку се сваком токену улазне секвенце додељује векторска репрезентација ознаке позиције у улазној секвенци (нпр. E_0, E_1, E_2, итд.). Збиром све три векторске репрезентације токена улазне секвенце се добија једна векторска

репрезентација токена која се прослеђује на улаз *BERT* модела. Обучавање *BERT* модела за задатак класификације се заснива на предвиђању излаза *[CLS]* токена који означава класу (нпр. да ли друга реченица улазног пара реченица чини наредну реченицу прве реченице улазног пара реченица).

Након обучавања, попут *ULMFiT* и *ELMo* језичких модела, *BERT* модел омогућава дообучавање за различите *NLP* задатке попут машинског превођења, аутоматског одговарања на питања, машинског закључивања, и тако даље. За сваки конкретни задатак се преузима обучени *BERT* модел који се фино подешава. Архитектура обученог *BERT* модела се минимално разликује од дообученог модела где се углавном мењају крајњи слојеви вештачке неуронске мреже. Пример обучавања и дообучавања модела аутора (Devlin *et al.*, 2018) је дат на Слика 9.



Слика 9 *Обучавања и дообучавање BERT модела*

BERT модел је постигао боље резултате од осталих актуелних модела на *GLUE* задацима (Wang *et al.*, 2018). Од *BERT* модела је настало много варијација модела базирано на промени корпуса, трајања обучавања и величине модела. Поред *ULMFiT*, *ELMo* и *BERT* модела су предложени и други модели који су обучавани на веома великим корпусима. Неки од њих су *Transformer-XL* (Dai *et al.*, 2019) и *XLNet* (Yang *et al.*, 2019) од стране компаније Гугл[7], затим *RoBERTa* (Liu *et al.*, 2019), *XLM* (Lample and Conneau, 2019) и *XLM-RoBERTa* (Conneau *et al.*, 2019) модели од стране компаније Фејсбук[8], и *GPT* (Radford *et al.*, 2018), *GPT-2* и *GPT-3* (Brown *et al.*, 2020) модели од стране *OpenAI*[9] компаније. Велики искорак у *NLP* области за мање познате језике (међу којима је и српски језик) представљају вишејезични трансформер модели који су обучени на више

[7] Компанија Гугл: https://about.google/
[8] Компанија Фејсбук: https://about.fb.com/company-info/
[9] Компанија *OpenAI*: https://openai.com/

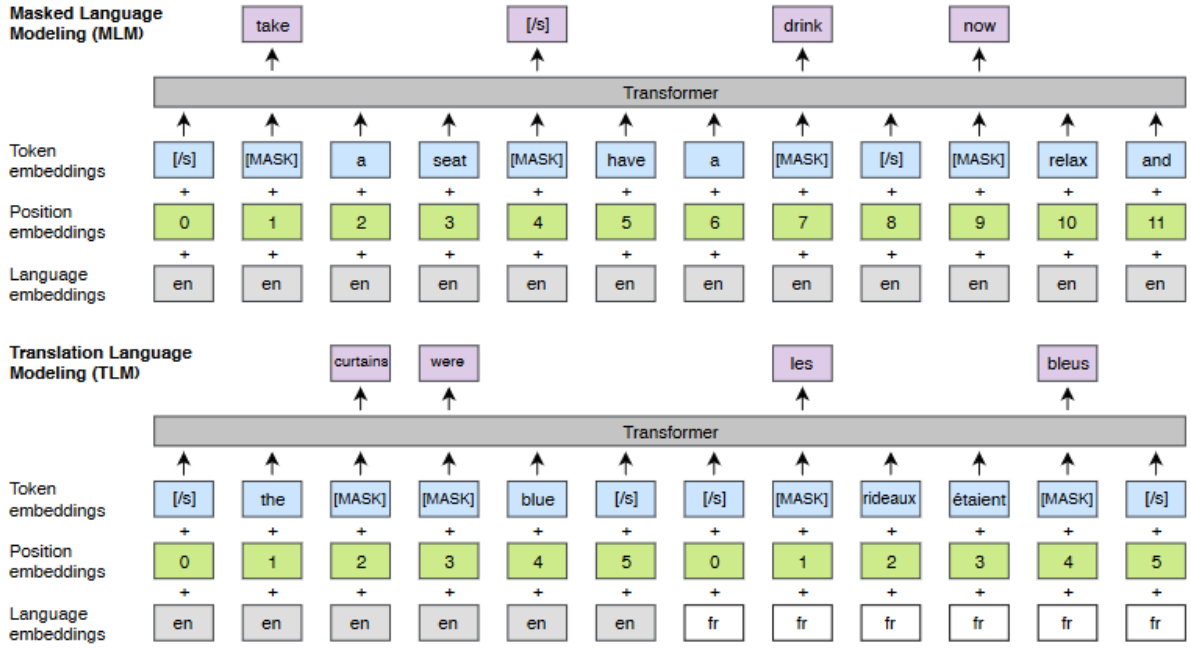
од 100 језика. Поред *BERT* модела, вишејезични модели који подржавају српски језик су *XLM* и *XLM-RoBERTa*, који ће бити детаљније описани у наставку.

Истраживачи компаније Фејсбук су развили вишејезични *XLM* модел (енгл. *Cross-lingual Language Model Pretraining - XLM*) који је базиран на трансформер архитектури. *XLM* модел представља побољшану верзију *BERT* модела којим су аутори остварили тада најбоље резултате за задатке класификације и машинског превођења. *XLM* модел користи другачију технику пред-процесирања улазних података и обучавање *BERT* модела врши по двојезичном механизму. Овај механизам представља обучавање модела са две секвенце речи за два језика у исто време како би модел научио релације између језика.

XLM модел при пред-процесирању улазних података користи технику компресије, односно енкодовања парова бајт вредностима (енгл. *Byte-Pair Encoding*). Ова техника енкодовања се базира на формирању дељеног речника између језика који се користи у обучавању. Речник се формира дељењем речи из улазних секвенци свих обучаваних језика на подречи, које се затим компресују. Алгоритам компресије се састоји у енкодовању најфреквентнијих симбола (карактера или низа карактера) једним бајт симболом. Алгоритам компресије се понавља док се не достигне дефинисан број итерација, односно степен компресије. Применом овог алгоритма компресије података се смањује величина дељеног речника која је неопходна за иновативни начин обучавања *BERT* модела.

Обучавање *XLM* модела се базира на комбинованом обучавању три језичка модела. Први модел је узрочни језички модел (енгл. *Causal Language Modeling - CLM*), заснован на трансформер архитектури, који се обучава са циљем моделовања вероватноће речи узимајући у обзир претходне речи у реченици. Други модел представља измењени *MLM* модел који се обучава на паралелним подацима тј. подацима сачињених од две секвенце на два језика. Речи улазне секвенце се трансформишу у векторске репрезентације применом формираног речника и сумирањем информација позиционог и језичког кодирања. Језичко кодирање додељује ознаку језика који се обучава и служи моделу да распознаје релације између сродних токена на различитим језицима. Овај надограђени *MLM* модел је означен од стране аутора као језички модел превођења (енгл. *Translation Language Modeling - TLM*). Трећи модел чини изворни *MLM* модел са другачијим узорковањем реченица у улазној секвенци и додатним језичким кодирањем. Аутори су експериментисали са комбинованим учењем *CLM*, *MLM* и *TLM* модела и закључили да повећању перформанси *XLM* модела највише доприноси обучавање *MLM* и *TLM* модела.

Пример обучавања *XLM* модела применом *MLM* и *TLM* модела аутора (Lample and Conneau, 2019) је дат на Слика 10.



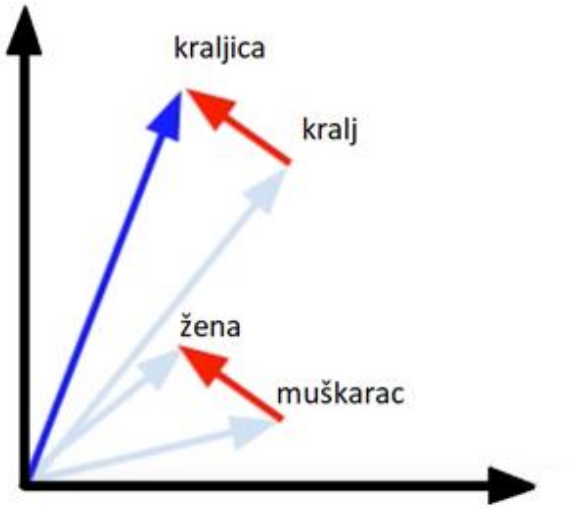
Слика 10 *Обучавање XLM модела*

Комплетни *XLM* модел, обученог применом *TLM* и *MLM* модела, је направио искорак у вишејезичном моделовању тј. обучавању модела на једном језику и примени на другом језику без додатних обучавајућиух података. *XLM* језички модел је надмашио перформансе *BERT* и претходних модела на задатку машинског превођења и генерално вишејезичном моделовању. Међутим, вишејезини *BERT* и *XLM* модели су ограничени на језике са више ресурса тј. већим корпусом.

Још већи успех у приступима вишејезичког моделовања је остварено *XLM-RoBERTa* моделом, који је представљен од стране компаније Фејсбук. *XLM-RoBERTa* модел је базиран на трансформер архитектури и представља надоградњу претходног *XLM* модела где је рутина обучавања преузета од *RoBERTa* модела, развијеног такође од компаније Фејсбук. Тачније, обучавање *XLM-RoBERTa* модела се базира на примени *MLM* модела на великој количини података, изостављајући *TLM* стратегију *XLM* модела.

Побољшање перформанси *XLM-RoBERTa* модела за језике са мало ресурса (нпр. српски језик) је остварено највише због огромног обучавајућег скупу података од два терабајта јавно доступних података за 100 језика. Применом Фејсбук интерног језичког модела у комбинацији са *fastText* моделом је додатно генерисано података, поготову за језике са мало ресурса. Аутори су истакли да је *XLM-RoBERTa* допринео у побољшању перформанси у односу на *BERT* и *XLM* моделе за задатке класификације, анотирања секвенци и аутоматског одговарања на питања.

Предности описаних језичких модела су висок квалитет векторске репрезентације речи добијених на основу велике количине података. Сличне речи у векторском простору се налазе близу једни других и речи могу имати више степена сличности. На основу дистанце или косинуса угла између вектора се могу закључити разне аналогије. На пример, реч *мушкарац* и реч *жена* се у векторском простору налазе близу један другом јер се појављују у истом контексту у тексту. Уколико је познат вектор речи *краљица*, на основу односа речи *мушкарц* и *жена* (разлике вектора) се може пронаћи вектор речи *краљ*, који је у истом односу са вектором речи *краљица*. Ове врсте аналогија у векторском простору се могу вршити и на типовима речи где, на пример, вектор глагола пливање се налази у близини вектора глагола трчање. Такође, односи познатих вектора главних градова и земаља се може применити и на остале векторе земаља или градова. Пример поређења вектора на основу међусобне аналогије је приказан на Слика 11.



Слика 11 *Аналогија речи у векторском простору*

2.2.1.3.2 Репрезентација секвенце речи

Репрезентација текстуалне секвенце сачињене од више од једне речи, попут документа, се може извршити на више начина. Најпознатији модел репрезентације докумената је врећа речи (енгл. *bag-of-words - BOW*) где се текстуална секвенца репрезентује као скуп речи (токена) и одређених одлика, без узимања у обзир граматику и редослед речи. Свака текстуална секвенца се токенизује и репрезентује као скуп јединствених речи који се назива речник термина. Тачније, формира се матрица термина за сваки документ (енгл. *term document matrix*), где су редови матрице

јединствени термини тј. речи из речника, а колоне матрице документи из корпуса. Вредности у матрици репрезентују одређену одлику попут фреквенције термина (енгл. *Term Frequency - TF*), односно број појављивања дате речи у датом тексту. Међутим, фреквенција термина није одлика којом се на најбољи начин репрезентују речи у тексту. Уобичајено најфреквентније речи у тексту су врсте предлога, везника и речца које, у зависности од типа *NLP* задатка, не носе корисне информације. Такође, високо фреквентне речи које се појављују у већини докумената у корпусу не носе корисне информације по којима би се документи могли међусобно поредити у векторском простору (нпр. косинус угла између вектора). Стога се фреквенцији термина додаје тежински фактор који представља инверзну фреквенцију термина у документу (енгл. *Term Frequency–Inverse Document Frequency - TF-IDF*).

Статистичка *TF-IDF* метода истиче важност термина, односно речи, у тексту (документу) у односу на цео корпус (Salton and Mcgill, 1986). Вредност *TF-IDF* се повећава бројем појављивања термина у документу а смањује бројем докумената у корпусу у којима се појављује дати термин. То значи да ће термини у документу који се често појављују и у осталим документима у корпусу имати малу *TF-IDF* вредност. Насупрот томе, термини у документу који се ретко јављају у осталим документима у корпусу имају велику *TF-IDF* вредност. Интиутивно, *TF-IDF* одређује колико је дата реч релевантна у одређеном документу. Формално гледано, одређивање *TF-IDF* вредности се може израчунати на следећи начин (Salton and Mcgill, 1986) . Уколико је дат корпус докумената *D* сачињен од докумената *d*, *TF-IDF* вредност за реч r из документа *d* се рачуна на следећи начин:

$$r_d = f_{r,d} * \log \frac{|D|}{f_{r,D}}$$

где је $f_{r,d}$ фреквенција речи r у документу *d*, $|D|$ величина корпуса докумената *D*, и $f_{r,D}$ фреквенција докумената корпуса *D* у којима се реч r појављује. За случај да је $f_{r,D}$ приближно једнако броју докумената у корпусу $|D|$, вредност логаритма је мала позитивна вредност у опсегу између нула и један, што значи да је вредност r_d мања од $f_{r,d}$ али пак позитивна вредност. Ово имплицира да је термин тј. реч r релативно честа у целом корпусу докумената *D* али ипак значајна за документ *d*. У супротном случају, да је $f_{r,D}$ веома мала позитивна вредност, вредност логаритма је већа од један, што говори да ће *TF-IDF* вредност r_d бити већа од $f_{r,d}$. Ово значи да је термин тј. реч r значајна за документ *d* и није честа у корпусу докумената *D*. У крајњем случају, да је $f_{r,D}$ потпуно једнака са бројем докумената у корпусу $|D|$, значи да се термин тј. реч појављује у сваком документу корпуса *D*, то доводи до тога да је $\log(1)$ једнак нули, те је стога и *TF-IDF* вредност нула. То говори да реч r не доноси значајне информације за било који документ корпуса *D*.

Међутим, *BOW* модел није погодан за *NLP* задатке где је битан редослед речи (нпр. идентификација аспеката и ентитета) па се као решење користи токенизација секвенци *n* узастопних речи (*n-gram* модел). Токени од једне речи се називају униграми (енгл. *unigrams*), токени од две узастопне речи (*2-grams*) се називају биграми (енгл. *bigrams*), триграми (енгл. *trigrams*) секвенце од три узастопне речи (*3-grams*), и тако редом. Токени са више од једне речи се често користе у конструкцији језичких особина када је потребно задржати информације о контексту. Резултат конструкције репрезентације текста засновнано на *BOW* и *n-gram* моделу је вектор нумеричких вредности које представљају неке одлике попут *TF* и *TF-IDF* вредности.

Векторској репрезентацији текста се могу придружити и додатне језичке особине из претходних корака пред-процесирања (Секције 2.2.1.1 и 2.2.1.2) попут мета особина (нпр. број речи, број карактера, број интерпункцијских знакова), *POS* тагова (нпр. да ли садржи именицу или придев), назива ентитета (нпр. да ли се јавља одређени тип ентитета), и тако даље. На овај начин се, у зависности од *NLP* задатка, могу побољшати репрезентације тескта.

2.3 Аспектно базирана сентимент анализа

Циљ *ABSA* јесте анализа мишљења и сентимента у односу на одређени аспект тј. циљани објекат. Стога се *ABSA* може дефинисати као уређена торка (*i*, *a*, *s*) где *i* представља извор тј. субјекат исказаног сентимента (нпр. особа, организација), *a* аспект тј. објекат на који се односи сентимент, и *s* тип сентимента (нпр. поларитет, емоција). У случају анализе сентимент поларитета на примеру студента „Предмет је веома занимљив али недостају материјали на нашем језику.“ се идентификују две торке са истим субјектом исказаног сентимента - студент. Додатно, једна торка садржи позитивни сентимент поларитет за аспект предмет, а друга негативни сентимент поларитет за аспект материјал. Уређеним торкама се може придружити и временска компонента на основу које ће се поредити промена сентимента за одређени аспект кроз време. Такође, сумирањем *ABSA* резултата у оквиру специфичног форума или анкете се може добити опште мишљење тј. одређује се да ли преовладава задовољство или незадовољство одређеним аспектом.

Према дефиницији аутора (Liu, 2015), аспекти су уобичајно именице и именичке фразе, али такође могу бити и глаголи, глаголске фразе, придеви, прилози, и друге језичке конструкције. Стога, аспекти се могу појављивати у експлицитном и имплицитном облику. Издвајање експлицитних аспеката подразумева анализу именица и именичких фраза које означавају објекат тј. носиоца мишљења. На пример, у реченици

„Овај филм је веома занимљив и поучан.“ именица *филм* представља експлицитни аспект који носи позитиван сентимент поларитет речи *занимљив* и *поучан*. Имплицитни аспекати се не појављују у форми именица и именичких фраза али указују на одређени аспект. На примеру домена рецензије филмова, у реченици „Глума је била маестрална и представљала истинско уживање.“ се аспект не појављује експлицитно али се на основу глагола *глума* може закључити да се ради о имплицитном аспкету глумца који у овом случају носи позитиван сентимент на основу речи *маестрална* и *уживање*. Међутим издвајање имплицитних аспеката је веома захтеван задатак и за човека да на основу глагола, глаголских фраза, придева или прилога препозна и одреди о ком аспкету се ради. На примеру из домена рецензија филмова, у реченици „Он је урадио добар посао.“ је тешко одредити да ли се позитиван сентимент поларитет речи *добар* односи на одређеног глумца, режисера или треће лице. У овом случају је исправније доделити општији аспект (или ентитет) који се односи на целу поставу филма, глумце, режисере, стилисте, итд.

Анализа мишљења на ниову аспеката се може извршити на нивоу целих реченица али је у случају више аспеката у реченици неопходно одредити позицију аспеката у реченици. Већи изазов представља одређивање који сентимент поларитет припада ком аспекту у реченици. Најједноставнији приступ је посматрање одређеног опсега најближих речи уоченог аспекта, где се у сентимент поларитет околних речи придодаје датом аспекту. Међутим, у случају дељења реченице на клаузе и фразе анализа се изводи на мањем сегменту текста, где се углавном налази један аспект. Овим приступом се отклања двоумљење око тога на који аспект се односи уочен сентимент поларитет. У случају да се не препозна ни један аспект посматраног домена, онда се у том случају подразумева општи или генерални аспект датог домена.

У *ABSA* области истраживања се појављује и случај апстракције носиоца мишљења где се прави разлика између аспекта и ентиета. Аспекти се у овом случају представљају као атрибути ентитета. На пример, реченица „Главни глумац филма ми се није свидео, али је филм веома интересантан.“ се под ентитетом сматра филм а глумац под његовим аспектом. На овај начин се анализа мишљења може посматрати из угла апстрактних ентитета или на нивоу аспеката. У литератури се јављају различити термини попут анализа ентитета (енгл. *entity-based sentiment analysis*), циљева (енгл. *target-based sentiment analysis*), тема (енгл. *topic-based sentiment analysis*), који се по методологији могу свести у један термин.

2.4 *Теоријске основе опште методологије за развој NLP система*

Решавање одређеног проблема применом *NLP* технике захтева дефинисан процес развоја модела и тестирања у циљу постизања што бољих перформанси на датом задатку. Тачније, развој *NLP* модела захтева методолошки приступ решавања датог проблема.

Почетни корак сваке методологије представља дефинисање циља тј. проблема који се решава. Уколико је проблем комплекснији тада се приступа дељењу проблема на специфичне задатке чијим решавањем се долази до решења. Затим се приступа дизајнирању решења где се *NLP* модели често посматрају као експерименти који се касније интегришу у интерфејс одређеног система. Сам процес формирања модела је уобичајно најдужа фаза у решавања одређеног проблема. Циљ је постићи велику тачност модела при решавању датог проблема. На успешност *NLP* модела, поред одабира алгоритма модела и начина обучавања, у великој мери утиче квалитет и количина података. Стога је потребно обратити велику пажњу на податке при дефинисању решења проблема.

Методолошки развој *NLP* модела представља цикличан процес развоја модела где се модел континуално развија и евалуира док се не постигну најбоље могуће перформансе. Пре самог циклуса развоја модела се врше припремни кораци формирања корпуса и пред-процесирање података. По завршетку развоја модела се врши тестирање и мерење крајњих перформанси модела. Графички приказ предложене методологије је дат на Слика 12.

Слика 12 *Методолошки развој NLP модела*

У првом кораку формирања корпуса се подразумева припрема података који ће се користити за решавање дефинисаног проблема. Анализирају се постојећи јавно доступни корпуси и врши њихова обрада и припрема података за примену одређеног *NLP* модела. У случају недостатка јавно доступних корпуса се приступа прикупљању података, пред-процесирању података (Секција 2.2.1) и анотирању истих. Анотирање подразумева структуирање података по дефинисаној шеми анотације (Секција 4.1). На крају пред-процесирања података се подаци деле на подскупове који се користе у наредним корацима развоја модела.

Подела корпуса се врши на скупове за обучавање, валидацију и тестирање модела. Величине скупова се најчешће узимају у односу 60-20-20 процената, где 60 процената чини обучавајући скуп, 20 процената валидациони и 20 процената тестни скуп података. У зависности од количине подататака, модела и задатка који се решава, димензије скупова података могу бити дефинисани у различитом односу. Уколико је скуп за обучавање веома мали онда се уместо валидационог скупа користи техника унакрсне валидације (енгл. *n-fold cross validation*), где се већи део обучавајућег скуп података користи за обучавање а остатак за валидацију. Дати процес се понавља одређен број пута и преузима се просечан резултат валидације. У случају класификације података на више класа је потребно обратити пажњу на равномерну расподелу података по класама на скупове за обучавање, валидацију и тестирање. Најчешће се подаци насумично измешају и одаберe дефинисан проценат података из сваке класе. На овај начин се спречава погрешна евалуација модела. У овом истраживању је корпус (Секција 4) подељен на подскупове за обучавање, валидацију и тестирање у односу 60-15-25 процената коришћењем насумичног избора базираном на униформној дистрибуцији.

У кораку формирања модела се врши одабир конкретног алгоритма или архитектуре модела. Подешавају се параметри модела, дефинише се процес обучавања, евалуације и оптимизације параметара модела. Овај корак има кључни утицај на успешност решавања дефинисаног проблема. Погрешним одабиром модела и начина обучавања и оптимизације параметара модела се и поред велике количине и квалитета података не могу постићи добри резултати.

Обучавање модела се врши применом обучавајућег скупа података. Мера квалитета формираног модела током обучавања се проверава у кораку валидације на скупу за валидацију модела. Уколико је процес обучавања успешно обављен онда формирани модел генерализује обучавајуће податке у довољној мери тако да се успешно примењује на скупу података за валидацију. Параметри модела се у оптимизују на основу резултата валидације у циљу боље генерализације података. Поред основних

параметара модела се оптимизују и такозвани хипер-параметри модела који су специфични за одређени тип модела (Секција 2.5.3.2.2.3.1).

На основу перформанси модела на валидационом скупу података се испитује задовољавање постављеног циља тј. решења проблема. Уколико модел не задовољава постављене циљеве онда се може приступити опционом кораку анализе грешака. У овом кораку се врши испитивање модела на основу грешака које је модел направио током примене на скупу за обучавање и валидацију. Циљ је открити узрок због ког модел прави одређене грешке и покушати изменити конфигурацију модела тако да их више не прави. Уколико је разлог количина или квалитет карактеристика података, онда се у наредном кораку врши измена и понавља цео претходни процес.

Када перформансе обученог модела на валидационом скупу задовољи постављени циљ, онда се квалитет обученог модела проверава у наредном кораку тестирања на до тада невиђеним подацима тј. тест скупу података. *NLP* модели се формирају на одређеном узорку података (корпусу) тако да се могу успешно применити у реалном сценарију када вредности података варирају. Стога се тестни скуп података не корсити у претходним корацима развоја модела и тиме се тестира понашање модела у реалном сценарију.

Перформансе модела се мере на основу излазних вредости модела. У случају класификације података, излазни подаци модела представљају класе које се пореде са улазним подацима и формирају се резултати који се интерпретирају на одређен начин. Постоји више различитих мера перформанси модела те ће се у наставку (Секција 2.4.1) описати најчешће коришћене мере класификације модела.

2.4.1 Евалуација модела *NLP* система

Постоје различити задаци који се обављају у оквиру *NLP* области као што су класификација текста, машинско превођење текста, сумирање текста и тако даље. За сваки *NLP* задатак се примењује специфичан начин евалуације, стога је фокус ове секције на задатку класификације текста. Предмет ове докторске дисертације је класификација текстуалних сегмената на класе сентимент поларитета и аспеката студирања.

Евалуација модела представља мерење перформанси модела. Подаци који се користе за обучавање и евалуацију класификационог модела се састоје од текстуалног сегмента и одговарајуће класе тј. лабеле. Током процеса анотирања (Секција 4.1) се врши додела класа сегментима текста од стране стручне особе (анотатора) из датог

домена. На основу анотираних података се може извршити аутоматска процена квалитета модела применом одређене методе евалуације.

Једна од најкоришћенијих метода евалуације класификационог модела подразумева мерење тачности (енгл. *accuracy*), прецизности (енгл. *precision*), одзива (енгл. *recall*) и Ф-мере (енгл. *F-measure*). Евалуација модела зависи и од броја класа, стога ће се прво изложити случај бинарне класификације података као што је класификација текста на позитиван и негативан сентимент поларитет. Други случај је задатак класификације података на више класа као што је аспектно базирана класификација, сентимент класификација (позитиван, негативан, неутралан), класификација емоција, и слично.

Уколико посматрамо случај бинарне класификације, подаци су означени једном од укупно две лабеле тј. класе (нпр. Класа0, Класа1). Процесом анотирања (Секција 4.1) су додељене стварне лабеле које се током еваулације узимају као тачне и пореде са предвиђеним лабелама које су добијене од одређеног модела. Дакле, циљ бинарне класификације је предвиђање лабеле Класа1 или Класа0 у зависности од фокуса модела. У овом случају ћемо размотрити предвиђање лабеле Класа1 на основу које се рачунају перформансе модела.

Прво се формира матрица конфузије (енгл. *confusion matrix*) која описује перформансе класификационог модела на посматраном скупу података за које су познате тачне вредности. Свака ћелија представља један од могућих исхода класификације. На главној дијагонали матрице конфузије се налазе суме исправно предвиђених класа примера. Тачно позитиван исход је број примера који су предвиђени класом Класа1 а чија стварна класа је исто Класа1. Обратно, тачно негативан исход број примера чија је предвиђена и стварна класа Класа0. На споредној дијагонали матрице конфузије се налазе суме погрешно предвиђених класа примера. Нетачно негативан исход је број примера чија је стварна класа Класа1 али је модел погрешно предвидео класу Класа0. У обратном случају, нетачно позитиван исход је број примера чија је стварна класа Класа0 али је погрешно предвиђена класа Класа1. Табела матрице конфузије (Табела II) је дата у наставку.

Табела II *Табела матрице конфузије*

	Предвиђена Класа1 (позитиван)	Предвиђена Класа0 (негативан)	
Стварна Класа1 (позитиван)	Тачно Позитиван (ТП)	Нетачно Негативан (НН)	Одзив
Стварна Класа0 (негативан)	Нетачно Позитиван (НП)	Тачно Негативан (ТН)	
	Прецизност		Тачност

Мера тачности представља проценат класа које је дати модел исправно предвидео (ТП и ТН) у односу на укупан број посматраних класа (ТП, НН, НП и ТН). Формалније, мера тачности се може представити у математичком облику на следећи начин:

$$\text{тачност} = \frac{\text{ТП} + \text{ТН}}{\text{ТП} + \text{НН} + \text{НП} + \text{ТН}}$$

Ова метрика тачности је довољна за случај да је број класа (Класа1, Класа0) равномеран, односно избалансиран у скупу података. Иначе, мера тачности би била погрешна процена модела без обзира на висок проценат тачности. На пример, ако скуп података садржи 10.000 коментара и од тога је само 100 позитивног сентимента а остатак негативног сентимента (9.900). Применом неког једноставнијег класификационог модела који ће, без обзира на улаз, увек предвиђати негативан сентимент излаз. То би значило да овај модел има прецизност од 99 процената ($\frac{9.900}{10.000} = 0{,}99$) што је очигледно погрешна процена квалитета модела. Из овог разлога неизбалансираног скупа података, што је чест случај у *NLP* области, се користе мере прецизности и одзива.

Посматрајмо репрезентативнији пример нешто сложенијег класификационог модела где небалансираност скупа података из претходног примера такође доводи до високог процента тачности (99 процената). Класификациони модел у овом случају успева да предвиди позитиван сентимент (Класа1) једном успешно и једном неуспешно.

У свим осталим случајевима предвиђа негативан сентимент (Класа0). Матрица конфузије у овом случају изгледа као у Табела III.

Табела III *Матрица конфузије за пример небалансираног скупа података*

	Предвиђена Класа1	Предвиђена Класа0
Стварна Класа1	ТП = 1	НН = 98
Стварна Класа0	НП = 1	ТН = 9.900

Мера прецизности представља проценат предвиђених класа које је дати модел предвидео исправно тј. класом Класа1. Од свих примера које је модел означио класом Класа1, колико је заправо примера анотирано класом Класа1. Формула за прецизност се рачуна на следећи начин:

$$\text{прецизност} = \frac{\text{ТП}}{\text{ТП} + \text{НП}}$$

Поред високе тачности модела (99 процената) у постављеном примеру прецизност износи 50 процената ($\frac{1}{2} = 0{,}50$), што реалније репрезентује дати случај. Дати модел је половично прецизан јер једну класу скоро увек успешно предвиђа а другу класу готово никад.

Мера одзива одражава проценат примера класе Класа1 који су испрвано предвиђене од стране модела. На примеру сентимент анализе, од свих примера позитивног сентимента (Класа1), колико је модел означио класом позитивног сентимента (Класа1). Формула за одзив се рачуна на следећи начин:

$$\text{одзив} = \frac{\text{ТП}}{\text{ТП} + \text{НН}}$$

За наведени пример одзив износи један проценат ($\frac{1}{99} = 0{,}01$) што указује да модел у један посто случајева успева да предвиди позитивну класу. Мере прецизности и одзива, за разлику од тачности, правилније одражавају перформансе модела у случајевима небалансираног скупа података. Постоји више метрика које укључују метрике прецизности и одзива у једну меру, од којих је Ф-мера највише коришћена у литератури.

Ф-мера репрезентује перформансу модела користећи мере прецизности и одзива. По својој дефиницији, Ф-мера садржи тежински фактор β којим се фаворизује тј. даје већа значајност прецизности или одзиву. Формула Ф-мере са тежинским фактором је дефинисана на следећи начин:

$$\text{Ф-мера} = \frac{(\beta^2 + 1) * \text{прецизност} * \text{одзив}}{\beta^2 * \text{прецизност} + \text{одзив}}$$

Међутим, највише је коришћена Ф-мера са јединичном вредношћу тежинског фактора ($\beta = 1$) чиме се утицај прецизности и одзива изједначава. Коначна формула Ф-мере са јединичним тежинским фактором је дефинисана на следећи начин:

$$\text{Ф-мера} = \frac{2 * \text{прецизност} * \text{одзив}}{\text{прецизност} + \text{одзив}}$$

На основу претходног примера и мера прецизности и одзива, Ф-мера износи 0,02 ($\frac{0,01}{0,51} = 0,02$) и означава веома слабу перформансу модела на датом небалансираном скупу података. Вредности Ф-мере су у опсегу од нула до један, где су боље перформансе модела кад је вредност ближа јединици.

За случај класификације података на више од две класе се разликује начин евалуације модела. Постоје два типа класификације на више класа где је главна разлика у броју предвиђених класа. Први тип класификације података на више класа представља случај више-лабеларне класификације (енгл. *multi-label classification*) где се један пример може означити са више од једне лабеле. Чешћи случај је други тип класификације на више класа где се један пример може означити само једном класом. Овај тип класификације је познат и под називом више-класне класификације где један пример може бити означен само једном класом. Матрица конфузије се у овом случају проширује на више класа и мере прецизности и одзива се рачунају засебно по класама. У табели која следује (Табела IV) је приказана матрица конфузије за случај више-класне класификације на три класе.

Табела IV *Табела матрице конфузије за случај класификације на више класа*

	Предвиђена Класа1	Предвиђена Класа2	Предвиђена Класа3	
Стварна Класа1	Тачно Класа1-Класа1 (ТК11)	Нетачно Класа2-Класа1 (НК21)	Нетачно Класа3-Класа1 (НК31)	Одзив Класа1
Стварна Класа2	Нетачно Класа1-Класа2 (НК12)	Тачно Класа2-Класа2 (ТК22)	Нетачно Класа3-Класа2 (НК32)	Одзив Класа2
Стварна Класа3	Нетачно Класа1-Класа3 (НК13)	Нетачно Класа2-Класа3 (НК23)	Тачно Класа3-Класа3 (ТК33)	Одзив Класа3
	Прецизност Класа1	Прецизност Класа2	Прецизност Класа3	Тачност

Рачунање мере прецизности, одзива и Ф-мере за сваку од класа је поступно идентично. Пример рачунања мере прецизности, одзива и Ф-мере за класу Класа1 је дат у наставку:

$$\text{прецизност}_{\text{к1}} = \frac{\text{ТК11}}{\text{ТК11} + \text{НК12} + \text{НК13}}$$

$$\text{одзив}_{\text{к1}} = \frac{\text{ТК11}}{\text{ТК11} + \text{НК21} + \text{НК31}}$$

$$\text{Ф-мера}_{\text{к1}} = \frac{2 * \text{прецизност}_{\text{к1}} * \text{одзив}_{\text{к1}}}{\text{прецизност}_{\text{к1}} + \text{одзив}_{\text{к1}}}$$

У случају потребе да се мере прецизности, одзива и Ф-мере изразе са једном мером по класи, онда се врши сумирање мера по класи и одређивање просека на три начина. Прво се формирају засебне бинарне матрице конфузије по класама, где је позитивна класа циљана класа а негативна класа сума осталих класа. Пример засебних бинарних матрица је дат у Табела V.

Табела V *Засебне матрице конфузије за случај класификације на више класа*

Класа1	Предвиђена Класа1	Предвиђена остале
Стварна Класа1	ТК11	НК21 + НК31
Стварна остале	НК12 + НК13	ТК22 + ТК33 + НК23 + НК32

Класа2	Предвиђена Класа2	Предвиђена остале
Стварна Класа2	ТК22	НК12 + НК32
Стварна остале	НК21 + НК23	ТК11+ ТК33 + НК13 + НК31

Класа3	Предвиђена Класа3	Предвиђена остале
Стварна Класа3	ТК33	НК13 + НК23
Стварна остале	НК31 + НК32	ТК11+ ТК22+ НК12 + НК21

Макро просек мере (енгл. *macro average*) подразумева рачунање мере засебно по класи, сумирање истих и одређивање просека по броју класа. На примеру прецизности, прво се израчунају прецизности по класама на основу засебних матрица конфузије (Табела V), затим се сумирају и поделе бројем класа (у овом примеру бројем три). Микро просек мере (енгл. *micro average*) подразумева формирање збирне матрице засебних матрица по класама (Табела VI) и на основу те матрице рачунати мере прецизности, одзива и Ф-мере. Макро тежински просек мере (енгл. *macro weighted average*) подразумева рачунање мере засебно по класи које се затим, пре сумирања, помноже тежинским фактором. Тежински фактор представља пропорцију сваке класе у скупу података (фреквенција класе подељена укупним бројем класа у скупу података).

У Табела VII је дат конкретан пример на основу ког је у наставку прецизније објашњено рачунање просека мере прецизности.

Табела VI *Збирна матрица засебних матрица конфузије из Табела V*

Збирна матрица	Предвиђене Класа1, Класа2 и Класа3	Предвиђена остале
Стварне Класа1, Класа2 и Класа3	ТК11+ТК22+ТК33	(НК21 + НК31) + (НК12 + НК32) + (НК13 + НК23)
Стварна остале	(НК12 + НК13) + (НК21 + НК23) + (НК31 + НК32)	(ТК22 + ТК33 + НК23 + НК32) + (ТК11+ ТК33 + НК13 + НК31) + (ТК11+ ТК22+ НК12 + НК21)

Табела VII *Конкретан пример засебних и збирне матрице конфузије за случај класификације на више класа*

Класа1	Предвиђена Класа1	Предвиђена остале
Стварна Класа1	10	10
Стварна остале	10	9.970

Класа2	Предвиђена Класа2	Предвиђена остале
Стварна Класа2	10	90
Стварна остале	10	9.890

Класа3	Предвиђена Класа3	Предвиђена остале
Стварна Класа3	90	10
Стварна остале	10	9.890

Збирна матрица	Предвиђене Класа1, Класа2 и Класа3	Предвиђена остале
Стварне Класа1, Класа2 и Класа3	110	110
Стварна остале	30	29.750

Макро просек мере прецизности се рачуна сумирањем засебних мера прецизности по класама и одређивање просека на основу броја класа. Дакле, засебне мере прецизности по класама на основу примера (Табела VII) су 0,50 ($\frac{10}{10+10} = 0{,}50$) за Класа1 и Класа2, и 0,90 ($\frac{90}{90+10} = 0{,}90$) за Класа3. Макро просек мере прецизности износи 63 процената ($\frac{0{,}50+0{,}90+0{,}50}{3} = 0{,}63$). Микро просек мере прецизности се рачуна на основу збирне матрице засебних матрица конфузије по класама (Табела VII). Микро просек мере прецизности износи 79 процената ($\frac{110}{110+30} = 0{,}79$). Макро тежински просек мере се рачуна множењем засебних мера прецизности са тежинским факторима. Уколико се посматра случај да су пропорције класе у скупу података тј. тежински фактори редом 0,10 ($\frac{1.000}{10.000} = 0{,}10$) за Класа1, 0,40 ($\frac{4.000}{10.000} = 0{,}40$) за Класа2, и 0,50 ($\frac{5.000}{10.000} = 0{,}50$) за Класа3. Онда макро тежински просек мере прецизности износи 70 процената ($0{,}50 * 0{,}10 + 0{,}50 * 0{,}40 + 0{,}90 * 0{,}50 = 0{,}70$).

Макро просек мере изједначава важност класа и свака мера класе подједнако доприноси просечној мери. Док микро мросек фаворизује фреквентније класе тим што имају већи утицај на просечну меру. Макро тежински просек мере равномерно подешава утицај засебних мера класа према фреквенцији датих класа у скупу података. У зависности од домена проблема, макро тежински просек мере може бити најпогоднији за изражавање перформанси модела класификације на више класа у случају неизбалансираности скупа података или када су феквентније класе веће важности од мање фреквентних (нпр. детектовања спам порука, медицинска предвиђања болести).

2.5 *Теоријске основе развоја система за анализу мишљења*

Након што су представљене опште теоријске основе везане за развој *NLP* система, у овој секцији су описане теоријске основе развоја система за анализу мишљења. Модели који се користе за развој система за анализу мишљења се могу груписати у четири категорије. То су модели засновани на речницима (енгл. *dictionary-based - DB*), модели засновани на правилима (енгл. *rule-based - RB*), модели машинског учења и хибридни модели (енгл. *hybrid*). Модели засновани на речницима и правилима су саставни део најстаријих традиционалних приступа у *NLP* области. Првобитно, речници су били формирани ручно а касније и аутоматски са неопходном евалуацијом стручњака из одређеног домена. Модели засновани на правилима користе унапред дефинисана правила попут регуларних израза и граматичких правила. Ови типови

модела су се показали као успешни у проналажењу шаблона (енгл. *pattern-matching*) али ограничени знањем дефинисаним правилима и речницима. Перформансе ових модела су мала прецизност и високи одзив, што значи да могу имати високе перформансе у специфичним случајевима а лошије у генералним случајевима. Из тог разлога се у новијим истраживањима модели засновани на речницима и правилима користе као помоћни део другог модела, формирајући хибридни модел.

У типичне приступе *NLP* области се убрајају модели засновани на машинском учењу попут машине потпорних вектора, Наивног Бајеса, стабла одлучивања (енгл. *decision trees*), логистичка регресија (енгл. *logistic regression*) и тако даље. Ови типови модела се ослањају на векторе језичких особина добијених пред-процесирањем (Секција 2.2.1) попут одређених речи, ентитета, типова и редоследа речи, множина речи, и тако даље. Ови типови модела су се показали као успешни у проналажењу апстрактнијих шаблона података.

Међутим, конструкција језичких особина је временски захтевна, зависно од конкретног *NLP* задатка и захтева доменско знање. Првобитно су језичке особине конструисане ручно и уз помоћ статистичких метода, након чега, су приступи дубоког учења уз велике корпусе олакшале и аутоматизовале процес конструкције језичких особина. Модели дубоког учења попут *ELMo* и *BERT* модела се убрајају у моделе засноване на машинском учењу. Ови модели самостално формирају векторе језичких особина (Секција 2.2.1.3) чиме се омогућило проналажење апстрактнијих језичких особина које не морају бити везане за конкретан *NLP* задатак. Такође, више типова алгоритама се могу комбиновати у један хибридни модел који ради као целина – на основу улаза се добија излаз као резултат целокупног модела.

У наставку су изложени модели који су коришћени у овој докторској дисертацији у сврси класификације, према горе описаним категоријама.

2.5.1 Модели засновани на речницима

Модели засновани на речницима припадају првобитним приступима истраживању *NLP* области. Речници се формирају према намени тј. задатку у *NLP* области на ком се примењују. У случају класификације текста, специфични речници садрже речи и фразе чије појављивање у тексту указује на припадност датог текста одређеној класи.

Речници се према начину формирања могу груписати у три групе (Секција 3.1). Првобитан приступ формирања је био ручан, напоран и временски захтеван приступ

који се касније углавном користи за проверу аутоматизованог приступа формирања речника. Другу групу чине приступи засновани на речницима који користе постојеће речнике синонима и антонима и технику генерисања речи (енгл. *bootstrap*) на основу речника синонима и антонима. Прво се ручно сакупља мали скуп речи који служи за претрагу постојећих језичких речника за њихове синониме и антониме (нпр. *WordNet*[6]). Трећу групу чине приступи засновани на корпусу који су мање ефикасни него приступи засновани на речницима. Примењује се у случају идентификације доменски специфичних сентимент речи и адаптације сентимент речника опште намене на одређени домен, користећи доменски корпус.

2.5.2 Модели засновани на правилима

Модели засновани на правилима припадају првобитним приступима истраживању *NLP* области. Правила у NLP области су доменски зависна и формирају се ручно од стране доменских стручњака. Процес формирања правила је временски захтеван и подложан грешкама људског фактора. Стога се модели засновани на правилима у данашње време користе као делови других модела. У наставку су описана два типа правила, правила заснована на регуларним изразима и граматичким правилима, којим се проналазе језички шаблони тј. фразе које се најчешће користе при коментарисању одређеног аспекта.

Регуларни израз (енгл. *regular expression*) представља формални језик који дефинише стандардну текстуалну синтаксу за репрезентацију шаблона претраге текста. Регуларни изрази омогућавају напредније претраживање текста од најједноставнијег алгоритма претраге стрингова, где се тражи егзакно поклапање претраживаног појма, до сложенијег алгоритма који дозвољава већу варијацију карактера. Постоји више стандарда којим се дефинише регуларни језик, стога ће се у наставку описати карактеристике регуларног језика *POSIX*[10] стандарда.

Поред алфанумеричких карактера у регуларним изразима се користе и мета карактери „[ ] ( ) { } ^ $. | * + \“ који имају своје специјално значење. Сваки карактер у регуларном изразу може имати буквално и специјално значење. Карактер *„A“* у регуларном изразу означава буквалну претрагу слова A, док се специјални израз *„[A-Z]“* односи на преграгу свих великих слова алфабета од слова A до слова Z. Из датог примера се може видети да регуларни изрази праве разлику између малих и великих слова.

[10] *POSIX* стандард: http://get.posixcertified.ieee.org/

Такође, за случај да је потребно претражити било који једноцифрен број онда се регуларни израз састоји из опсега цифара *„[0-9]“*.

Угласте заграде *„[ ]“* у регуларним изразима означавају дисјункцију карактера који се налазе између угластих заграда. То значи да се речи попут *nastava*, *nastavi* и *nastavu* могу претражити регуларним изразом попут *„nastav[aiu]“*. Уколико је потребно обрнута логика од наведене, односно да је потребно претражити све речи које почињу са *nastav* а не завршавају се на слова *a*, *i* и *u* (нпр. *nastavnica*), онда се користи мета карактер ^ у регуларном изразу *„nastav[^aiu]“*. Мета карактер „^“ се користи и за означавање почетка линије или стринга. За примену дисјункције карактера на већој секвенци карактера се користе мета карактери заграда *„( )*“. На пример, претрага речи *nastavnica* и *nastava* се може постићи регуларним изразима попут *nastavnica*|*nastava*, *nastav(nica*|*a)* и *nastav(nic)?a*.

Мета карактер знака питања *„?*“ замењује један или ниједан карактер или опсег карактера које следује у регуларном изразу. На пример, регуларни израз *„nastavnik?*“ дефинише шаблон за проналажење речи попут *nastavnik* и *nastavni*, јер се слово *k* може појавити једном или ни једном. Поменути мета карактер се може применити и на опсег карактера, те се у регуларном изразу *„nastavni[a-z]?*“ односи на било који карактер у опсегу од слова *a* и *z*. За претрагу израза који се разликују у једном карактеру (осим карактера за нови ред) на одређеном месту се користи мета карактер „.“ (тачка) у регуларном изразу. На пример, речи *imali u imati* се могу претражити регуларним изразом “*ima.i*”.

У случају претраге понављајућих карактера у регуларном изразу се користе *„+*“ и *„**“ мета карактери. Мета карактер *„**“ замењује ниједан или више карактера или низа карактера које следује у регуларном изразу. Регуларни израз *„xy*z*“ претражује речи попут *„xz*“, *„xyz*“, *„xyyz*“ и *„xyyyyyyz*“. Применом поменутог мета карактера на више карактера као у регуларном изразу „[xyz]*“ се односи на речи попут *„x*“, „y“, „z“, „xz“, *„zyx*“ и тако даље. У случају да је потребно ограничити број понављања карактера *x* онда се примењује регуларни израз *„x{a,b}*“ којим се ограничава минимум појављивања на *a* и не више од *b* пута. Мета карактер *„+*“ замењује један или више карактера или низа карактера које следује у регуларном изразу. Регуларни израз „3+“ претражује све једноцифрене и вишецифрене бројеве које почињу са бар једном цифром 3, дакле „3“, „332“, „333“ и тако даље. За претрагу било ког броја, једноцифреног или вишецифреног, онда се користи опсег цифара у регуларном изразу *„[0-9]+*“.

За претрагу мета карактера у тексту се користи мета карактер „\“. Регуларним изразом *„\[(.+)\]*“ се угласте заграде не посматрају се као мета карактери са специјалним значењем него се претражује појављивање више узастопних карактера између угластих заграда. Такође, мета карактер „\“ се користи у скопу специјалних карактера, на пример,

„\w“ за издвајање карактера речи, „\w*“ за издвајање речи, „\d“ за издвајање цифара, „\b“ за означавање почетка или завршетка речи, „\s“ за издвајање „белих“ тј. невидљивих знакова (размака, табова, карактера за нови ред). За претрагу карактера који се налазе на самом крају стринга онда се користи мета карактер „$“.

Регуларни изрази се могу обогатити и граматичким правилима како би се претраживали комплекснији изрази. Граматичка правила се односе на редослед врста речи у изразима који се претражују. На пример, регуларним изразима се могу дефинисати шаблони којим се претражују речи у различитим међусобним односима попут придев-именица, прилог-именица-глагол, предлог-именица, придев-придев-именица, и тако даље. Врсте речи се могу добити применом алата за граматичко означавање (Секција 2.2.1.2) или ручно формираним речницима придева, прилог, именица и глагола.

На примеру регуларних израза, правило које издваја граматички шаблон придев-именица се дефинише као „((fenomenal|najbolj|pozitiv|izuzet|dobar|raspoloz |...)\w*)[ ]((nastavnik|covek|legend|osob|…)\w*)“. Група придева и група именица је одвојена размаком („[ ]“) чиме се покрива шаблон од две речи. Уколико је потребно уопштити ово правило тако да се једна реч може наћи између групе придева и именице, онда се може додати следећи регуларни израз „(\w+)?[ ]“. Низ тачака „...“ у регуларном изразу не представља део регуларног израза него мења остале придеве из речника придева и именица. Правила заснована на регуларним изразима се могу писати применом *JAPE*[11] и *MIXUP*[12] библиотека.

2.5.3 Модели машинског учења

Класични модели машинског учења су веома ефикасни у генерализацији знања из података и примењивању знања над новим подацима у различите сврхе тј. задатке. Уколико одређени пример податка сличан неком примеру који је модел сусрео током учења тј. обучавања, модел употребљава знање стечено учењем како би проценио тражени излаз. Циљ је да се формира систем где ће се модел континуално побољшавати на одређеном задатку. У наставку су описане две групе модела машинског учења, група

[11] *GATE* оквир за писање *JAPE* правила: https://gate.ac.uk/
[12] MinorThird оквир за писање *MIXUP* правила
http://gnteam.cs.manchester.ac.uk/wiki/index.php?n=Resources.MinorThird

класичних модела машинског учења и група модела вештачких неуронских мрежа и дубоког учења.

2.5.3.1 Класични модели машинског учења

У наставку су описани најчешће коришћени модели машинског учења заснованим на ручно формираним језичким особинама – модел Наивни Бајес, модел к најближих суседа и модел машине потпорних вектора.

2.5.3.1.1 Модел Наивни Бајес

Модел Наивног Бајеса је базиран на Бајесовој теореми, односно теорији вероватноће, која податаке посматра као строго независне. Уколико појаву података посматрамо као догађаје A и B, Бајесова теорема се може изразити следећом формулом:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

Условна вероватноћа $p(A|B)$ се тумачи као вероватноћа да ће се десити догађај A уколико се десио догађај B, и обратно за $p(B|A)$. Условна вероватноћа $p(A|B)$ се рачуна као производ условне вероватноће $p(B|A)$ и маргиналне вероватноће догађаја A $p(A)$, подељених са маргиналном вероватноћом догађаја B $p(B)$. Маргиналана вероватноћа догађаја B $p(B)$ се не мења без обзира на догађај A, те се стога може занемарити. Сада се формула рачунања условне вероватноће $p(A|B)$ може поједноставити на следећи начин:

$$p(A|B) = p(B|A)p(A)$$

Уколико се догађај B престави као вектор карактеристика x_i, где је $i \in N$, онда се претходна формула рачунања условне вероватноће $p(A|B)$ може преформулисати на следећи начин.

$$p(A|x_1, x_2, \dots, x_n) = p(x_1, x_2, \dots, x_n|A)p(A)$$

Додатно, формула условне вероватноће $p(A|B)$ се може разложити применом заједничког модела вероватноће и правила уланчавања:

$$p(x_1, x_2, \dots, x_n|A)p(A) = p(x_1, x_2, \dots, x_n, A)$$
$$= p(x_1|x_2, \dots, x_n, A)\, p(x_2|x_3, \dots, x_n, A) \cdots p(x_{n-1}|x_n, A)p(x_n|A)p(A)$$

За сваку кобинацију карактеристика x_i је потребно много параметара модела за учење и огроман скуп података. Стога се у овом случају приступа претпоставкама које поједностављују посматрани случај. Претпоставком да су карактеристике x_i међусобно независне у односу на догађај *A*, условне вероватноће добијене применом правила уланчавања се могу заменити на следећи начин:

$$p(x_i|x_{i+1}, \dots, x_n, A) = p(x_i|A)$$

На основу чега се вероватноће у заједничком моделу вероватноћа могу „наивно“ помножити на следећи начин:

$$p(x_1, x_2, \dots, x_n|A)p(A) = p(x_1, x_2, \dots, x_n, A) =$$

$$p(x_1|A)\, p(x_2|A) \dots p(x_n|A)p(A) = \; p(A) \prod_{i=1}^{n} p(x_i|A)$$

У контексту класификације текста овакав модел се користи са наивном претпоставком да се појава једне речи у документу не зависи од појаве друге речи у документу. Класификација текстуалних докумената и њених дискретних језичких особина се врши моделима који су засновани на мултиноминалној (енгл. *multinomial distribution*) и Берноулијевој (енгл. *Bernoulli distribution*) дистрибуцији особина. Берноулијев *NB* модел је погодан за бинарну класификацију докумената где се за одлике вектора језичких особина користе бинарне вредности, односно да ли је у улазном податку присутна одређена особина или не. Ако су одлике вектора језичких особина представљење фреквенцијом појављивања термина (*BOW* језичких особина), онда је за овај случај погоднији мултиноминални *NB* модел.

Ако је документ *d* представљен вектором језичких особина f_i, где је $i \in N$, мултиноминални *NB* (*MNB*) класификатор предвиђа једну класу $\hat{k}$ од скупа дискретних класа $k \in K$ која има највећу вероватноћу у датом документу. Одабир класе која има највећу вероватноћу се добија максималне постериори методом (argmax у формули). Излазна вредност $\hat{k}$ класификационог модела заснованом на Бајесовој теореми се може израчунати на следећи начин:

$$\hat{k} = \underset{k \in \{1,2,\dots,K\}}{\operatorname{argmax}}\; p(k) \prod_{i=1}^{n} p(f_i|k)$$

Излаз мултиноминалног *NB* класификатора се често мапира у логаритамски простор чиме постаје линеарни класификатор. Стога, коначна формула за рачунање излазне вредности $\hat{k}$ је дата у наставку:

$$\hat{k} = \underset{k \in \{1,2,\ldots,K\}}{\operatorname{argmax}} \log p(k) + \sum_{i=1}^{n} \log p(f_i|k)$$

Упркос овако наивном приступу класификатори Наивног Бајеса су се показали успешним у неким реалним ситуацијама (нпр. детекција спам порука). Предност ових модела је та да им није потребно пуно обучавајућих података да би извршили класификацију.

2.5.3.1.2 Модел к најближих суседа

Класификација података се код модела к најближих суседа не заснива на формирању модела (попут *NB* модела) него на једноставном чувању инстанци из обучавајућег скупа података (енгл. *instance-based learning*).

Процес класификације се извршава простим већинским гласањем к најближих суседа (енгл. *k-Nearest Neighbors - k-NN*). Модел *k-NN* складишти обучавајуће податке као векторе карактеристика и њихових класа у вишедимензионалном векторском простору. Складиштење обучавајућег података заправо представља процес учења *k-NN* модела. Податку који се класификује се додељује класа која има највише представника дате класе у њеном окружењу.

Постоје две имплементације *k-NN* модела у зависности од начина посматрања суседних података. У првој имплементацији овог типа модела се вредност к узима као вредност са покретним зарезом која представља радијус. На основу радијуса се посматрају најближи суседи податка који се класификује. Ова имплементација модела је погодна за неуниформну расподелу података мање димензионалности. У другој имплементацији вредност к је целобројна и представља број најближих суседа који ће учествовати у гласању, односно чија класа ће бити узета у обзир при додели класе податку који се класификује. Повећањем вредности к се повећава зависност података, односно повећава се број података чија класа се разматра што може утицати на смањење ефекта шума (подаци који су изузеци од већине других из дате класе). Додатно се уводи тежински фактор којим се даје већи значај суседима који су ближи податку који се класификује. Најчешће коришћа метрика сличности два податка (документа) је Еуклидско растојање (Danielsson, 1980) које је погодно за континуалне

вредности, док су за дискретне вредности попут текста погодије метрике Хаминг растојања (Norouzi *et al.*, 2012) и сличности на основу косинуса угла два вектора. У случају не дефинисања тежинског фактора свих к суседа има исту тежину при гласању додељивања класе.

Формалније, процес класификације *k-NN* модела се према ауторима (Wang and Zhao, 2012) може дефинисати на следећи начин. Како би се класификовао документ d_x, *k-NN* модел рангира документе $d_t \in D$ из векторског простора докумената обучавајућег скуп D. Затим се користе класе (лабеле) од к најближих суседних докумената за предвиђање класе документа d_x. Класама ових докумената се додаје тежински фактор на основу одређене мере сличности. У овом случају ће бити размотрена косинусна мера сличности два вектора тј. документа d_x и d_t. Косинус два вектора документа се се може извести из унутрашњег производа (енгл. *inner product*) два вектора на следећи начин:

$$d_x \cdot d_t = \|d_x\|\|d_t\| \cos(\theta)$$

$$\cos(\theta) = \frac{d_x \cdot d_t}{\|d_x\|\|d_t\|}$$

Ако је сваки документ *d* репрезентован вектором карактеристика $f_1, f_2, \dots, f_n$ и једном класом $c_t \in C$ од скупа дискретних класа C, онда се косинусна сличност два вектора $sim(d_x, d_t)$ може написати на следећи начин:

$$sim(d_x, d_t) = \cos(\theta) = \frac{\sum_{i=1}^{n} f_{xi} \times f_{ti}}{\sqrt{\sum_{i=1}^{n} f_{xi}^2} \sqrt{\sum_{i=1}^{n} f_{ti}^2}}$$

Где је бројилац разломка сума производа компоненти тј. карактеристика вектора d_x и d_t, а делилац разломка производ еуклидове норме вектора тј. дужина вектора d_x и d_t.

Након израчунатих сличности докумената, се врши класификација тј. одређивање којој класи d_x припада. Сличности документа d_x и осталих докумената из обучавајућег скуп $d_t \in D$ ($sim(d_x, d_t)$) се сортирају у опадајућем редоследу (прво су документи са највећом сличношћу) и врши се сумирање сличности првих к суседа по класама. За сваку класу $c_t \in C$ се сумира к првих сличности на следећи начин:

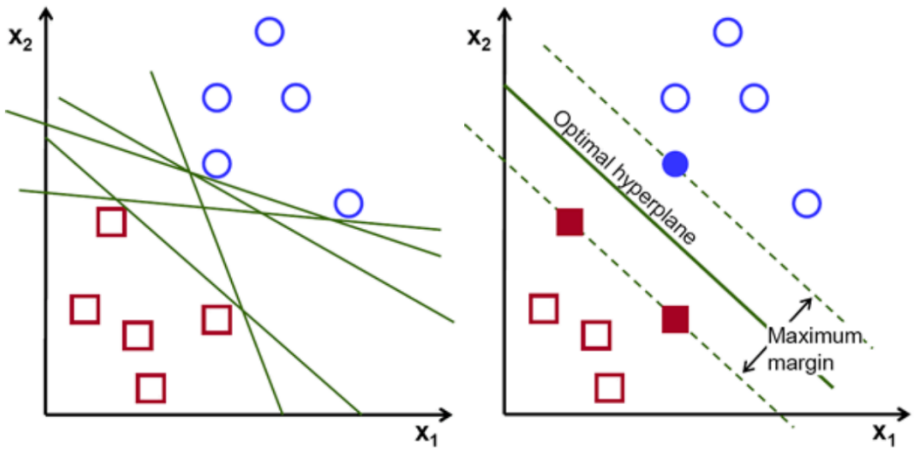
$$p(d_x, c_t) = \sum_{i=1}^{k} sim(d_x, d_{ti}) P(d_{ti}, c_t)$$

Где је $P(d_{ti}, c_t)$ бинарна вредност која означава да ли дати документ d_{ti} припада циљаној класи c_t. На основу сумираних вредности сличности по класама $p(d_x, c_t)$ за документ d_x се врши додела класе са највећом сумом сличности.

Предност *k-NN* модела је једноставаност за имплементацију и обучавање јер нема потребу за подешавањем параметара модела. Поред класификације, користи се и за претрагу сличних докумената попут система препоруке књига, филмова, докумената, и тако даље. Међутим, због начина учења модела, порастом броја каркатеристика докумената *k-NN* модел постаје спорији и мање ефикасан по меморију.

2.5.3.1.3 Модел машине потпориних вектора

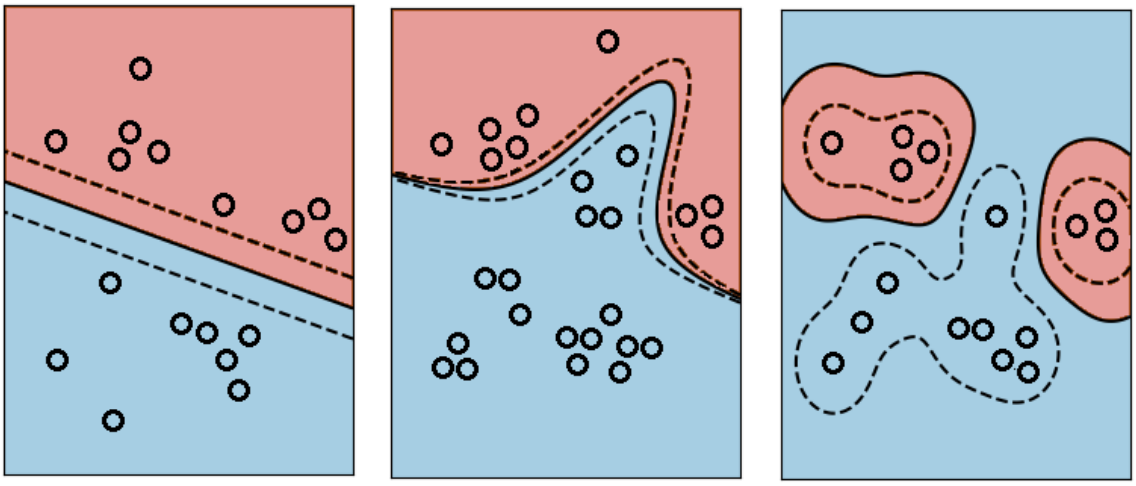
Модел машине потпорних вектора је у основи бинарни класификатор који податке репрезентује као тачке у вишедимензионалном простору (векторе) и за циљ има проналажење вишедимензионалне хиперравни (енгл. *hyperplane*) која најбоље раздваја ове тачке. Оптимална вишедимензионална хипер раван је она која прави највеће раздвајање тј. маргину између две класе. Одређивање оптималне хиперравни је базирано на анализи најближих чланова група тачака, такозваних потпорних вектора. Нови подаци се мапирају у исти векторски простор и врши класификација на основу стране хиперравни где налазе. Пример могућих хиперравни у дводимензионалном простору и одабир најбоље је дат на Слика 13. Потпорни вектори који утичу на одређивање оптималне хиперравни се налазе на испрекиданим линијама (Слика 13, десно).



Слика 13 *Пример SVM модела у дводимензионалном простору*

Уколико је хипер раван $n-1$ димензије у n димензионалном векторском простору, онда се овај модел назива линеарни *SVM* класификатор. То значи да је функција хиперравни базирана на линеарној комбинацији карактеристика. У дводимензионалном векторском простору хипер раван преставља праву линију којом се раздвајају подаци који су линеарно раздвојиви (Слика 13). Међутим, *SVM* модел може

вршити и нелинеарну класификацију применом различитих метода језгра (енгл. *kernel methods*). Методе језгра мапирају улазне податке у високодимензионални векторски простор (скаларни производ вектора) чиме је линеарном класификатору омогућено да научи нелинеарну функцију. Пример метода језгра за линеарну, полиномијалну и радијалну функцију је приказан на Слика 14.



Слика 14 *Пример различитх метода језгра*

Иако је *SVM* модел бинарни класификатор, постоје два начина на који се могу применити у случају вишекласне класификације. Први начин је да се обучава асамбл тј. више бинарних класификатора по стратегији „један против једног“ (енгл. *one-versus-one*). То подразумева да за сваки пар класа постоји један бинарни класификатор (нпр. за четири класе има шест бинарних класификатора). Класификација се врши на основу гласања већине (енгл. *max-wins*) асамбла где се податку додељује она класа која има највише гласова тј. предвиђања. Други начин подразумева обучавање класификатора по стратегији „један против осталих“ (енгл. *one-vs-rest*). У овом случају се за n класа формира n бинарних класификатора при чему су за негативне класе узете све преостале класе. На пример, за случај класификације на четири класе, када је циљана прва класа, за негативну класу су узете преостале три класе. Класификација се врши на основу гласања највеће вредности (енгл. *winner-takes-all*) асамбла где се податку додељује она класа чији бинарни класификатор је предвидео највећу континуалну вредност (попут вероватноће). Иако *SVM* модел није базиран на вероватноћи, захваљујући додатном скалирању (Platt, 1999) се бинарни излаз се може трансформисати у вероватноће.

Формално гледано, *SVM* класификациони модел са линеарним језгром се може објаснити на следећи начин. Уколико имамо обучавајући скуп података $D(x, y)$ који је сачињен од парова $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ улазних x_i и излазних вредности $y_i \in \{-1,1\}$. Ради лакшег објашњења модела се за негативну класу бинарног класификатора узима негативна вредност јединице. У домену анализе текста, улазне

податке x_i представљају документи а излазне податке y_i класе. Сваки документ x_i је представљен вектором карактеристика $f_1, f_2, \dots, f_n$ димензије n. Циљ *SVM* модела је пронаћи хиперраван максималне маргине која дели групу докумената x_i чија је класа $y_i = 1$ и групу докумената x_i чија је класа $y_i = -1$. Једначина хиперравни се може формулисати као скаларни производ ортогоналног вектора хиперравни w и вектора карактеристика докумената x (енгл. *dot product*) на следећи начин:

$$w \cdot x_i - b = 0$$

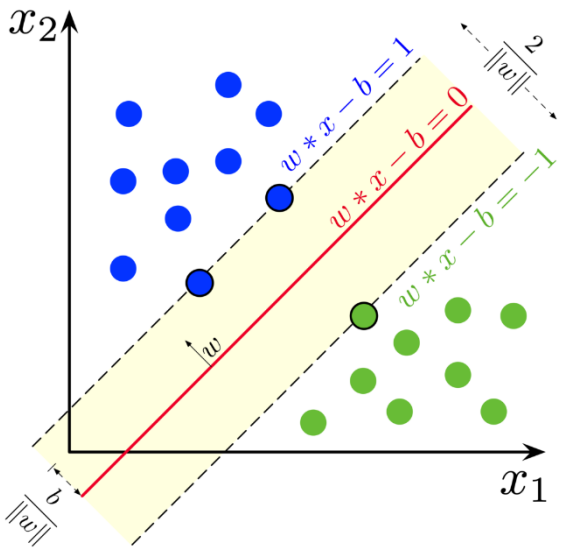
Где параметар b представља корекциони елемент једначине хиперравни. Ортогонални вектор хиперравни w може а не мора бити јединичан (магнитуда је један). Оваква једначина хиперравни ће довести до тога да било која хеперраван која раздваја две групе докумената је подједнако добра (Слика 13, лево). Међутим, *SVM* модел тежи ка постизању веће сигурности, односно ка оној хиперравни која има највећу маргину раздвајања између две групе докумената (Слика 13, десно). Да би се постигао овај циљ, додају се још два додатна услова. Поред услова хиперравни додатно се посматрају услови маргине односно две паралелне хиперравни које граниче маргину. Ове две хиперравни се ослањају на потпорне векторе класа, односно најближе документе класа хиперравни. У аналогији са вредностима класа ($y_i \in \{-1,1\}$), једначине ових додатних хиперравни се могу дефинисати на следећи начин:

$$w \cdot x_i - b = 1$$

$$w \cdot x_i - b = -1$$

Документи који се у векторском простору налазе изнад $w \cdot x_i - b = 1$ хиперравни припадају групи докумената чија је класа $y_i = 1$, док документи који се налазе испод $w \cdot x_i - b = -1$ хиперравни припадају групи докумената чија је класа $y_i = -1$. На овај начин се постављају услови да се вектори докумената морају налазити на исправној страни маргине.

Ако графички посматрамо дати проблем (Слика 15), хипрераван у дводимензионалном простору је приказана црвеном бојом. Два услова маргине, односно две паралене хиперравни, су приказани испрекиданом линијом. Вредност $\frac{b}{\|w\|}$ одређује ширину размака између хиперравни и групе вектора докумената у смеру нормале хиперравни. Док вредност $\frac{2}{\|w\|}$ одређује ширину маргине односно размак између две групе докумената. Дакле циљ *SVM* модела је да вредност $\frac{2}{\|w\|}$ буде максимална, односно да вредност $\|w\|$ буде минимална.



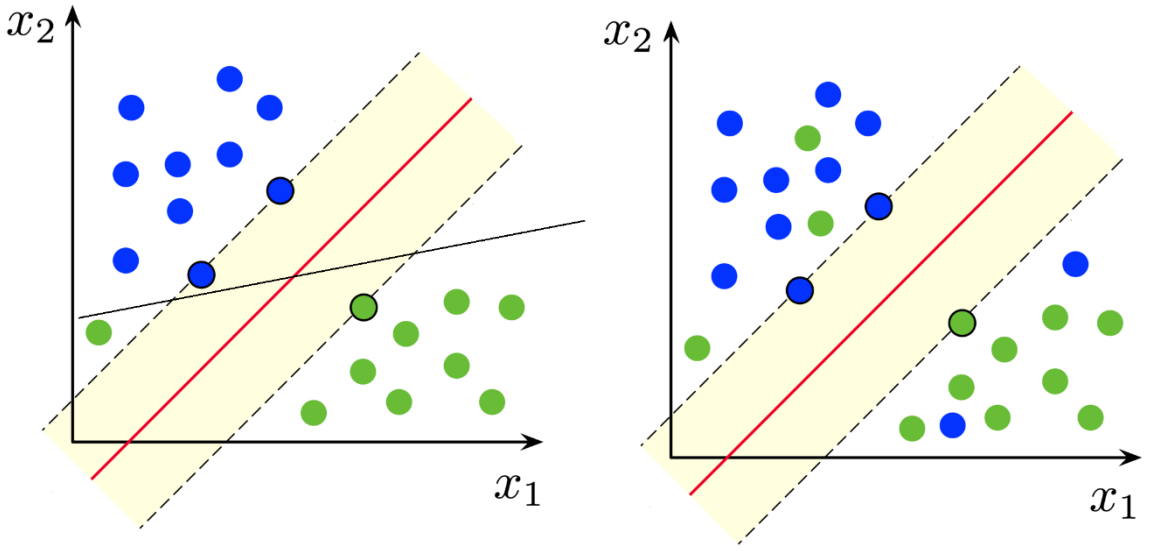
Слика 15 *Одређивање хиперравни SVM модела*

Проблем тражења хиперравни се може представити као проблем оптимизације функције хиперравни. Ако се излаз *SVM* модела представи ознаком $\hat{y}_i$ онда се услови маргине хиперравни могу репрезентовати једном неједначином на следећи начин:

$$\hat{y}_i = f(x_i) = w \cdot x_i - b$$

$$y_i \cdot \hat{y}_i \geq 1$$

Циљ оптимизационог проблема је минимизовати вредност $\|w\|$ тако да наведена неједначина буде задовољена. Овим се проналази хипераван која има стриктну маргину која потпуно раздваја векторски простор две класе докумената. Међутим, чест случај је када проблем бинарне класификације није могуће решити стриктном маргином због одступајућих вредности (енгл. *outliers*). Одступајуће вредности су у овом случају они документи који се у векторском простору налазе ван своје групе (заједнице) и често међу документима друге класе. Због одступајућих вредности проблем класификације постаје нелинеаран и потребно је изменити начин проналажења хиперравни или применити други метод језгра На примеру овакве једне ситуације (Слика 16) се може видети да у неким случајевима проблем није нелинеарне природе (Слика 16, лево) и да се може пронаћи хипераван али без строге маргине (црна линија), што је циљ *SVM* модела. Међутим, у другом случају (Слика 16, десно), проблем је нелинеарне природе и није могуће пронаћи оптималну хипераван са строгом маргином без примене друге методе језгра.



Слика 16 *Одступајуће вредности и проблем нелинеарности*

Међутим, овај случај нелинеарне природе се може решити релаксирањем маргине *SVM* модела тако да се толерише одређен број одступајућих вредности. То значи да се за оптималну хиперраван може изабрати она са највећом маргином (Слика 16, црвена линија) без обзира на поједине одступајуће вредности. У овом случају се претходна оптимизациона функција замени неком другом функцијом грешке. У проблему оптимизације се функција грешке користи као услов ка коме се тежи током учења модела. Углавном се у функцији грешке тежи ка постизању глобалног минимума (у случају да постоји више локалних минимума). Модел који не успева да испуни задати услов добија грешку те се стога оптимизациона функција и назива функцијом грешке. У случају да модел током учења испуни задати услов, функција грешке враћа нулту вредност. Постоји више типова функција грешке које се примењују у различитим случајевима, где ће се у Секцији 2.5.3.2.2.2 описати најкоришћеније. За проблем класификације са максималном маргином се најчешће користи функција грешке линеарне зависности (енгл. *hinge loss function*). У том случају, оптимизациона функција *SVM* модела L се може дефинисати на следећи начин:

$$L = \frac{1}{n}\sum_{i=1}^{n} max(0, 1 - y_i \cdot f(x_i;\, w_i)) + \gamma \frac{1}{2n}\sum_{i=1}^{n} w_i^2$$

Главни циљ оптимизационе функције је минимизовати функцију грешке L. Први део оптимизационе функције представља просечну грешку свих података (докумената) обучавајућег скупа, применом функције грешке линеарне зависности. Уколико је неједначина хиперравни ($y_i \cdot \hat{y}_i \geq 1$) задовољена, тј. предвиђена класа $\hat{y}_i$ и тачна класа y_i су истог знака, онда функција грешке враћа нулту вредност. У супротном случају, када неједначина хиперравни није задовољена, онда је повратна вредност функције грешке

пропорцијална дистанци датог вектора документа x_i од хиперравни. Други део оптимизационе функције ($\gamma \frac{1}{2n} \sum_{i=1}^{n} w_i^2$) чини регуларизациони део којим се релаксира маргина *SVM* модела. Тачније, регуларизационим делом једначине се врши кажњавање великих вредности w коефицијената (повећањем грешке за дате коефицијенте) чиме се спречава случај прекомерног обучавања модела (Секција 2.5.3.2.2.4). Вредношћу параметра γ се одређује утицај регуларизације на w коефицијенате, односно колико је маргина стриктна. За мале вредности параметра γ се постиже стриктна маргина, стање као и без регуларизационог дела оптимизационе функције. Већим вредностима параметра γ се повећава вредност грешке за w коефицијенте, затим се оптимизацијом минимизује утицај w коефицијената и тиме постиже релаксиранија маргина (енгл. *soft margin*). Сувише великим вредностима параметра γ доводи до подобучавања модела (Секција 2.5.3.2.2.4). У литератури се често овај проблем посматра из друге перспективе и уводи се параметар C којим се регулише релаксираност маргине. Оптимизациона функција у том случају се дефинише на следећи начин:

$$L = C \sum_{i=1}^{n} max(0, 1 - y_i \cdot f(x_i;\, w_i)) + \frac{1}{2} \sum_{i=1}^{n} w_i^2$$

Вредност параметра C је обрнуто сразмерна вредности парамтра γ ($C = \frac{1}{\gamma}$). За много велике вредности C се постиже стриктна маргина хиперравни *SVM* модела, док се мањим вредностима релаксира маргина хиперравни и дозвољава присуство одступајућих вредности (Слика 16). На овај начин се *SVM* моделу са линеарним језгром омогућава класификација података линеарне природе који су због одступајућих вредности постали проблем нелинеарне природе. У случају да овај приступ не даје добре перформансе се приступа промени методе језгра.

SVM модел је ефективан са високодимензионалним простором и захваљујући кернелима је њихова примена широка. У случају када димензионалност прелази број примера (дужина вектора карактеристика је веће димензије од скупа података), тада је *SVM* модел мање ефективан. Тада је избор кернела и термина регуларизације (Секција 2.5.3.2.2.4.1) неизоставан део за избегавање прекомерног обучавања модела (Секција 2.5.3.2.2.4). Интеграцијом *SVM* модела са векторском репрезентацијом података се добија модел који се показао као један од најбољих за задатак класификације текста.

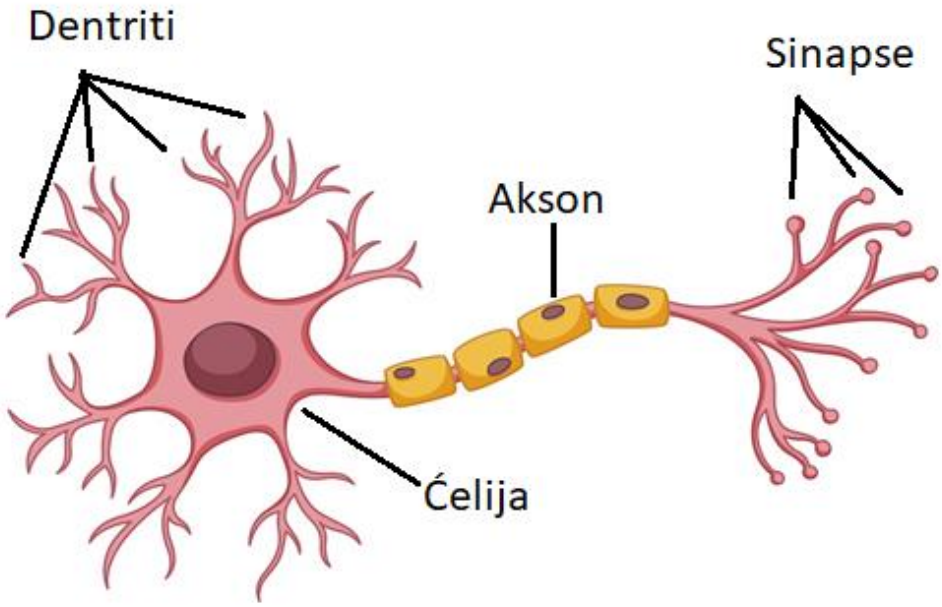
2.5.3.2 Модели вештачких неуронских мрежа и дубоког учења

Иако се машинско учење може окарактерисати као учење предвиђања на основу претходног искуства, приступ учења заснованом на вештачким неуронским мрежама (енгл. *Artificial Neural Network - ANN*) се базира на комплексној репрезентацији података. Захваљујући структури и нелинеарности модела вештачких неуронских мрежа се могу решити различити комплексни задаци. У Секцији 2.2.1.3 су представљени приступи формирања векторске репрезентације речи који су засновани на вештачким неуронским мрежама. Стога ће у овој секцији бити изложени детаљи вештачке неуронске мреже.

У наставку ће бити описане вештачке неуронске мреже са фокусом на *NLP* проблем класификације. Од специфичких типова вештачких неуронских мрежа ће бити описане конволуционе вештачке неруонске мреже (енгл. *Convolutional Neural Network - CNN*) и рекурентне вештачке неуронске мреже (енгл. *Recurrent Neural Network - RNN*), чија примена у *NLP* области је забележила велики пораст. Након тога ће бити описан процес учења и проблем оптимизације модела. На самом крају ће бити описане архитектуре модела дубоког учења који су коришћени у овом истраживању.

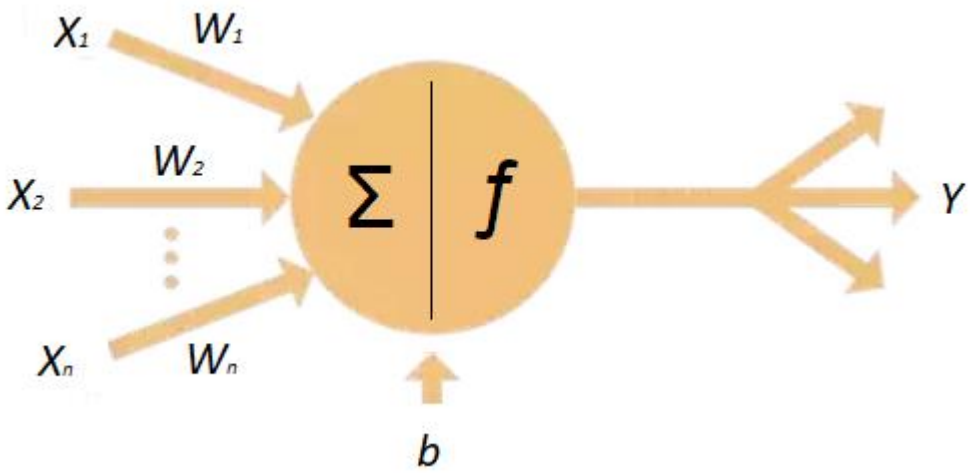
2.5.3.2.1 Вештачке неуронске мреже

Модел вештачких неуронских мрежа је настао по угледу на мрежу неурона која се налази у људском мозгу. Основна јединица вештачке неронске мреже је вештачки неурон који представља поједностављен математички модел биолошког неурона. Структура биолошког неурона (Слика 17) се у основи састоји од ћелије, дентрита (енгл. *dendrites*), аксона (енгл. *axon*) и синапси (енгл. *synapses*). Основна јединица неруона је ћелија која кроз дентрите прима улазне сигнале, обрађује их и затим путем аксона и контактних тачака тј. синапси преноси на друге неуроне.



Слика 17 *Структура биолошког неурона*

На сличан начин се може математички репрезентовати структура вештачког неурона (Слика 18). Улазни подаци вештачког неурона (дентрити) представљају излазе других вештачких неурона (аксони) који су повезани синапсама. Синапсе у математичком моделу представљају тежине, односно јачине колико одређени улазни сигнал има утицај на посматрану ћелију. Одређивање значаја улазних сигнала за дату ћелију се назива знањем које се огледа у подешавању (учењу) тежина синапси. У телу ћелије се врши математичко сумирање производа улазних сигнала са тежинама синапси, чему се додаје бијас параметар (енгл. *bias*). Бијас параметар не зависи од улаза неруона и служи за додатно прилагођавање излаза. Након тога се претходни резултат пропушта кроз функцију активације (енгл. *activation function*) која одређује да ли ће доћи до активирања тј. прослеђивања сигнала наредном вештачком неурону. Излаз ћелије односно вештачког неурона је сигнал (аксон) који се даље преноси другим вештачким неуронима.



Слика 18 *Структура вештачког неурона*

Формално, ако су улазни сигнали вештачког неурона означени са $x_1, x_2, \dots, x_n$, тежине синапси са $w_1, w_2, \dots, w_n$, бијас параметар са b, онда се излаз вештачког неурона z може дефинисати на следећи начин:

$$z = \left(\sum_{i=1}^{n} w_i x_i \right) + b$$

Другим речима, сума ($\sum_{i=1}^{n} w_i x_i$) представља скаларни производ вектора w и x ($w \cdot x$). Приказана формула представља дефиницију алгоритма бинарне (линеарне) класификације, познатијег под називом перцептрон (егнл. *perceptron*), коришћеног и у Секцији 2.5.3.1.3. Нелинеарност вештачког неурона се постиже пропуштањем вредности z кроз функцију активације f на следећи начин:

$$y = f(z)$$

Постоји више типова функција активације (Секција 2.5.3.2.2.1) које се користе у различитим случајевима и не мора бити иста код свих вештачких неурона. Уколико се за функцију активације узме, на пример, сигмоидна функција (енгл. *sigmoid function*) онда вештачки неурон представља једноставни бинарни класификатор тј. модел логистичке регресије (Pregibon, 1981).

Скуп вештачких неурона, организованих у слојеве, који су на одређени начин повезани међу собом чине целину тј. модел вештачке неуронске мреже. За вектор улазних вредности се произведе једна или више излазних вредности (у зависности од *NLP* задатка). Начин на који су вештачки неурони организовани и како се врши обрада (израчунавање) излазних вредности представља архитектуру вештачке неуронске мреже. Међу најједноставнијим и најпознатијим архитектурама вештачке неуронске

мреже је архитектура вишеслојног перцептрона (енгл. *multilayer perceptron*). У овој архитектури су вештачки нурони организовани у слојеве који су поређани у ред (секвенцу), где су свака два суседна слоја међусобно потпуно повезана.

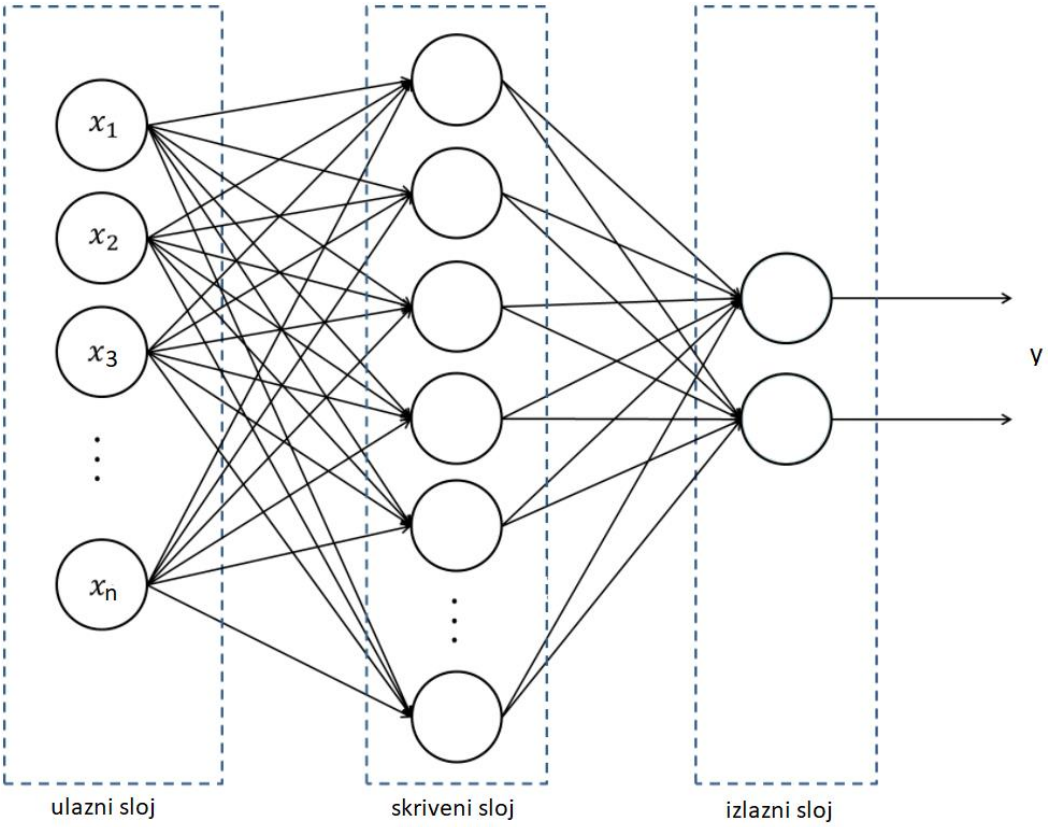
Алгоритам по којем се одвија учење вештачке неуронске мреже се базира на следећим корацима. Улазни подаци вештачке неуронске мреже се итеративно пропагирају кроз слојеве до излаза вештачке неуронске мреже (егнл. *feed forward*). У излазном слоју се добијају вредности које представљају предвиђање вештачке неуронске мреже. Добијене вредности предвиђања се пореде са очекиваним вредностима и рачуна грешка предвиђања применом функције грешке (енгл. *loss function*)[13]. Пропагирањем вредности грешке итеративно у назад, од излазног слоја до улазног слоја (егнл. *feed backward*), представља процес учења вештачке неуронске мреже када се подешавају синаптичке тежине између вештачких неурона. Примена наведених корака алгоритма на свим подацима обучавајућег скупа представља једну епоху (егнл. *epochs*) обучавања. Процес учења се одвија кроз више епоха докле год се не оствари најмања могућа грешка предвиђања. Овде може доћи до проблема прекомерног и недовољног обучавања модела (Секција 2.5.3.2.2.4).

Битан корак пре почетка обучавања вештачке неуронске мреже је постављање иницијалних вредности синаптичких тежина. При томе је потребно избећи проблем симетрије, односно случај када се вештачки неурони у скривеном слоју понашају исто (симетрично) тј. дају исте излазе и тиме не садрже „ново“ знање. У том случјау иницијализација синаптичких тежина нултим вредностима није решење него се примењује насумичан одабир иницијалних вредности или применом одређене функције расподеле попут нормалне расподеле.

Графички гледано, вештачка неуронска мрежа архитектуре вишеслојног перцептрона се може представити као граф чији чворови су вештачки неурони који су међусобно, између слојева, потпуно повезани. Улазни подаци вештачке неуронске мреже се представљају улазним слојем чворова. Колико има карактеристика улазног податка (нпр. речи у вектору) толико ће улазни слој имати вештачких неурона. Излазни слој чини један или више вештачких неурона тј. чворова чије излазне вредности представљају предвиђања модела. Уколико је у излазном слоју присутан један вештачки неурон онда се ради о бинарној класификацији, док се за случај вишекласне класификације користи више од једног вештачког неурона у излазном слоју. Слој између улазног и излазног слоја се назива скривени слој. Додавање више скривених слојева

[13] У литератури на енглеском језику се помиње и термин функције трошка (енгл. *cost function*) који се односи на грешку у целом обучавајућем скупу података. Другим речима, функција трошка рачуна просек грешака појединачних обучавајућих података добијених функцијом грешке.

омогућава проналажење, односно израчуанвање, комплекснијих карактеристика вештачке неуронске мреже. Пример овог типа архитектуре вештачке неуронске мреже је дат на Слика 19.



Слика 19 *Архитектура вишеслојног перцептрона*

2.5.3.2.2 Проблем обучавања и оптимизације

Процес обучавања вештачке неуронске мреже представља измену параметара модела тежина синапси и бијаса тако да модел производи очекивана предвиђања. Дакле, подаци за обучавање се доведу на улаз вештачке неуронске мреже и пропусте кроз мрежу. У зависности од иницијално подешених тежина синапси и типа функција активације у вештачким неуронима се производе различите трансформације података. Након тога се на основу излаза вештачке неуронске мреже (предвиђања) израчуна грешка предвиђања применом одређене функције грешке. Применом одређеног алгоритма оптимизације и функције грешке се рачунају промене излаза за дати улаз (градијенти) на основу којих се врше измене синаптичких тежина у вештачкој неуронској

мрежи. Цео процес представља једну епоху и понавља се до кад перформансе модела не буду задовољавајуће.

Повећањем дужине трајања процеса обучавања тј. броја епоха се повећава комплексност модела који не даје пропорцијално боље перформансе модела. У зависности од више параметара модел може бити сувише добро прилагођен подацима за обучавање (преобучен) тако да неуспешно врши класификације над новим подацима. Насупрот томе, уколико модел није добро прилагођен подацима за обучавање онда није довољно обучен (подобучен) да даје очекивана предвиђања ни за обучавајуће ни за тестне податке. Стога процес обучавања представља један комплексан процес оптимизације модела у који су укључени многи параметри које је потребно подесити на прави начин. Кључни кораци процеса обучавања и оптимизације модела вештачке неуронске мреже су описани у наставку.

2.5.3.2.2.1 Функције активације

Предност вештачких неуронских мрежа је заснована на њиховој могућности да одређени проблем реше применом скупа нелинеарних функција. Вештачки неурони у оквру свог израчунавања примењују одређену нелинеарну функцију чији излаз одређује шта ће наредни вештачки неурон или слој вештачких неурона добити на улаз. Скуп нелинеарних функција, у оквиру вештачких неурона, формира моћан математички модел који је у стању да научи комплексне карактеристике у подацима. У зависности од излазне вредности вештачког неурона надовезујући вештачки нурон се може активирати или остати неактиван (тежине синапси постају занемарљиво мале). Стога се функције које се примењују у вештачком неурону називају функцијама активације. Постоји много врста функција активације те ће се стога у наставку описати само најкоришћеније.

2.5.3.2.2.1.1 Сигмоидална функција

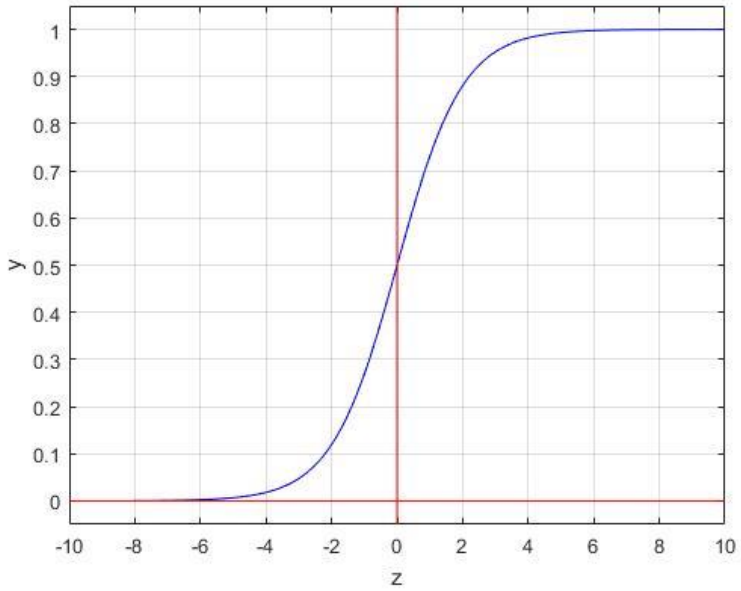
Једна од најкоришћенијих и најпознатијих функција активације је сигмоидална функција. Сигмоидална функција је име добила по изгледу графичке репрезентације на латинично слово *s* (Слика 20) која се такође назива и логистичком функцијом (енгл. *logistic function*). Једначина излаза бинарног класификатора, односно перцептрона, се

може представити збиром скаларног производа тежина синапси w и улазних вредности x и бијас параметра b на следећи начин:

$$z = w \cdot x + b$$

Где је x вектор улазних вредности вештачког неурона, w вектор тежина синапси вештачког неурона и b вредност бијас параметра. Излаз бинарног класификатора је бинарна вредност – нула или јединица. У случају да је излаз потребно представити вероватноћом тј. вредностима у опсегу $(0, 1)$, онда се користи сигмоидална функција. Графичка репрезентација сигмоидалне функције је приказана на Слика 20 а формула у наставку:

$$y = \sigma(z) = \frac{1}{1 + e^{-z}}$$



Слика 20 *Сигмоидална функција*

Предност сигмоидалне функције је то што реалне вредности мапира у опсег између нуле и јединице, што је и циљ за вероватноћу. Множењем излазне вредности сигмоидалне функције са вредошћу 100 се добија проценат вероватноће. Због своје скоро линеарности у вредности нуле и јединице ова функција све велике вредности попут одступајућих вредности мапира на вредности близу нуле или јединице. Међутим, скоро линеарни делови сигмоидалне функције представљајуману за дубоке вештачке неуронске мреже.

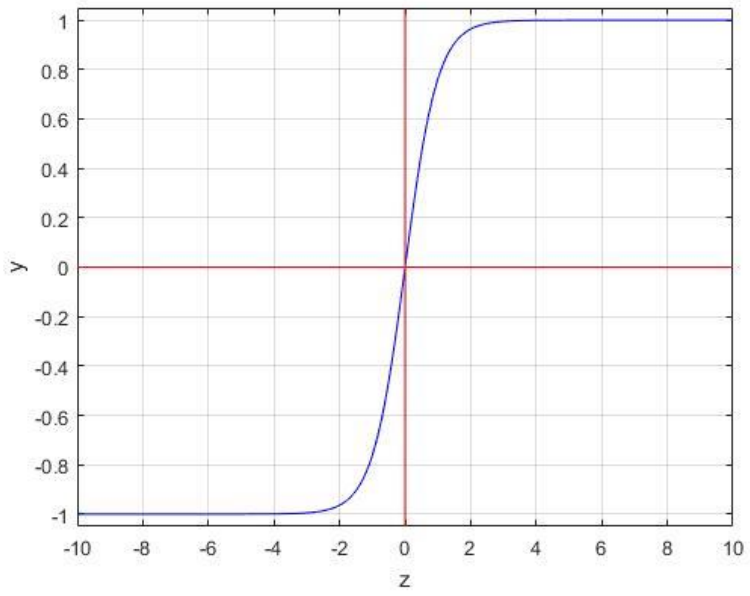
Велике промене на улазу сигмоидалне функције ће узроковати малим променама на излазу у скоро линеарним деловима функције. То значи да ће градијент

тј. извод сигмоидалне функције за много велике и много мале вредности улаза бити веома мали, близу нуле. Мале вредности градијента у дубоким вештачким неуронским мрежама доводи до нестајућег градијента (енгл. *vanishing gradient*), односно застоја у обучавању мреже.

2.5.3.2.2.1.2 Функција хиперболичне тангенте

Постоји више варијација сигмоидалне функције од којих је и функција хиперболичне тангенте (енгл. *hyperbolic tangent function*). Она представља једну варијанту сигмоидалне функције чија излазна вредност је у опсегу $(-1, 1)$. Графичка репрезентација сигмоидалне функције је приказана на Слика 21 а формула у наставку:

$$y = \sigma(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$



Слика 21 *Функција хиперболичне тангенте*

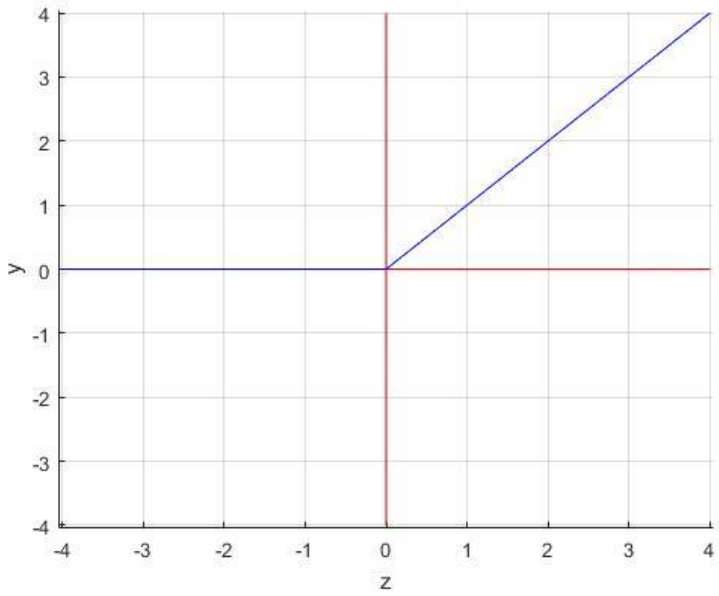
Предност фунције хиперболичне тангенте је што задржава негативни предзнак у подацима тиме што негативне вредности мапирати у опсег $(-1, 0)$. Позитивне вредности мапира у опсегу $(0, 1)$. Вреднсти близу нуле ће остати у датом опсегу. Мана функције хиперболичне тангенте у скоро линеарним деловима функције је иста као код

сигмоидалне функције и односи се на нестајући градијент у дубоким вештачким неуронским мрежама.

2.5.3.2.2.1.3 Функција исправљене линеарне јединице

Функција која ради делом на принципу линеарне функције ($f(x) = x$) где су негативне вредности мапиране на нулу се назива исправљена линеарна јединица (енгл. *Rectified Linear Unit - ReLU*). То значи да су улазне вредности веће од нуле пропуштене на излаз а негативне вредности поравнате на нулу. Излазне вредности *ReLU* функције су у опсегу $(0, \infty)$. *ReLU* функција је блиска по начину функционисања биолошког неурона. Убрзава учење вештачке неуронске мреже тиме што спречава нестајање градијента. Графичка репрезентација *ReLU* функције је приказана на Слика 22 а формула у наставку:

$$y = \text{relu}(z) = \max_{z}(0, z)$$



Слика 22 *Функција исправљене линеарне јединице - ReLU*

ReLU функција има и својих различитих варијанти које користе у одређеним случајевима. Једна од мана *ReLU* функције је и нестајање неурона услед дуже неактивности. Негативне вредности на улазу *ReLU* функције доводе до нултих вредности градијената, чиме тежине одређених неурона остају непромењене. Последично, свака наредна вредност ће посредством истих тежина бити трансформисана у негативну вредност што ће довести до застоја у обучавању. Због овог случаја је предложена

пропустљива *ReLU* функција (енгл. *Leaky ReLU function*) која за негативне вредности додаје линеарни део функције. Уколико се опсег пропустљивости негативних вредности параметризује, онда се добија параметризована *ReLU* функција (енгл. *Parametric ReLU function*). Формула параметризоване *ReLU* функције је дата у наставку:

$$y = \text{param\_relu}(a, z) = \begin{cases} z, & z > 0 \\ a \cdot z, & z \leq 0 \end{cases}$$

У зависности од вредности параметра a се дефинише опсег мапирања негативних врендости. Уколико је вредност параметра a једнака 0,01 онда се ради о пропустљивој *ReLU* функцији. На овај начин се омогућава пропуштање малих вредности градијената када вештачки неурон није активан. Излазне вредности попустљиве *ReLU* функције су у опсегу $(-\infty, \infty)$.

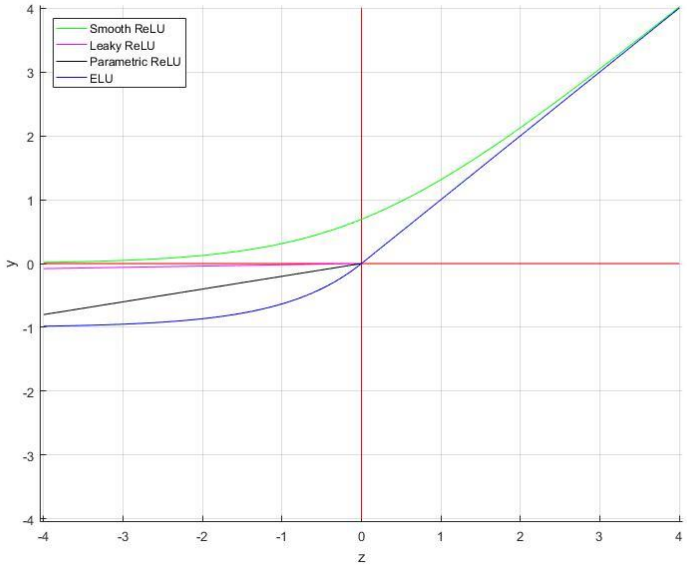
Такође, једна од мана *ReLU* функције је непостојање извода у нули па се стога користи углађена *ReLU* функција (енгл. *Smooth ReLU, softplus, function*). Ова функција ублажава линеарност и искључивост негативних вредности у пределу нуле по функцији природног логаритма. Излазне вредности су истом опсегу $(0, \infty)$ као и код *ReLU* функције. Формула углађене *ReLU* функције је дата у наставку:

$$y = \text{soft\_relu}(z) = \ln(1 + e^z)$$

Још једна варијанта *ReLU* функције која превазилази обе поменуте мане *ReLU* функције је функција експоненцијалне линеарне јединице (енгл. *Exponential Linear Unit* - *ELU*). Попут попустљиве и параметризоване *ReLU* функције, експоненцијални део *ELU* функције превазилази проблем нестајућих неурона за негативне вредности. Такође, попут углађене *ReLU* функције, *ELU* функција ублажава линеарност око вредности нуле и тиме превазилази проблем непостојања извода у нули. Излазне вредности *ELU* функције су у опсегу $(-\infty, \infty)$. Формула *ELU* функције је дата у наставку:

$$y = \text{elu}(z) = \begin{cases} z, & z \geq 0 \\ e^z - 1, & z < 0 \end{cases}$$

Графичка репрезентација варијација *ReLU* функција, углађене, параметризоване и пропустљиве *ReLU* функције, и *ELU* функције су приказане на Слика 23.

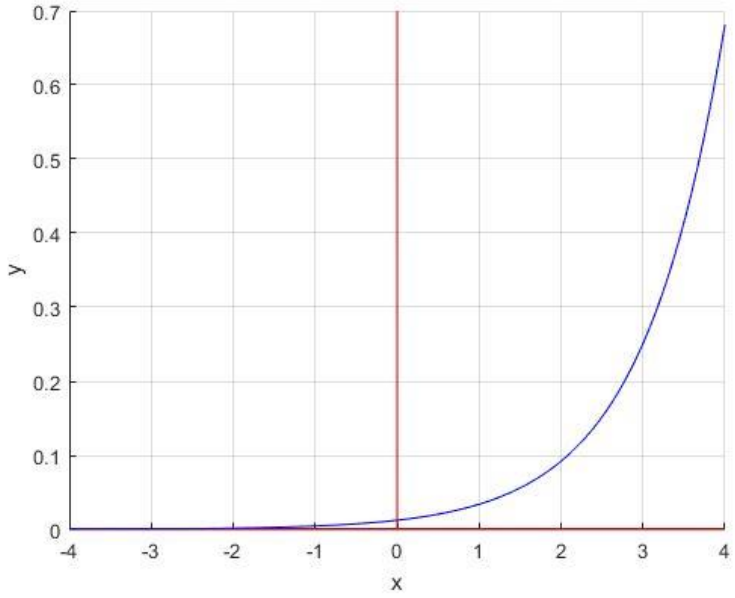


Слика 23 *Варијације ReLU функције*

2.5.3.2.2.1.4 Софтмакс функција

У случају да је у излазном слоју вештачке неуронске мреже потребно различите излазне вредности вештачких неурона нормализовати на опсег вероватноће $(0, 1)$ онда се користи софтмакс функција (енгл. *softmax function*). То значи да ће се сви улази вештачког неурона x мапирати на опсег вероватноћа $(0, 1)$ на тај начин да је укупна сума излазних вредности једнака јединици. Улазне вредности су нормализоване пропуштањем кроз експоненцијалну функцију. Графичка репрезентација софтмакс функције је приказана на Слика 24 а формула у наставку:

$$y = \text{softmax}(x) = \frac{e^x}{\sum_{i=1}^{n} e^{x_i}}$$



Слика 24 *Софтмакс функција*

2.5.3.2.2.2 Функције грешке

У процесу учења модела вештачке неуронске мреже је потребна мера колико је модел тренутно успешан тј. колико су вредности предвиђања модела $\hat{y}$ близу стварних вредности y добијених процесом анотације. Циљ процеса учења је подесити параметре модела тако да предвиђања модела $\hat{y}$ буду ближе тачним вредностима y што је више могуће. Разлика предвиђене вредности $\hat{y}$ и стварне вредности y се назива грешком. Грешка се рачуна применом неке од функција грешке, које се још називају функцијама губитка или трошка (Секција 2.5.3.2.1). На основу функције грешке се врши оптимизација модела тако да грешка предикције буде што мање.

Најједноставнија функција грешке је функција квадратне грешке ($\frac{1}{2}(\hat{y}-y)^2$) (енгл. *Mean Squared Error* - *MSE*) која, међутим, није погодна јер садржи локалне минимуме који отежавају рад оптимизационим алгоритмима (Секција 2.5.3.2.2.3). У наставку су описана функција грешке које су коришћене у овом истраживању. Прво је изложена функција грешке линеарне зависности која је коришћена код модела машине потпорних вектора (Секција 2.5.3.1.3), затим функција грешке унакрсне ентропије (енгл. *cross-entropy loss function*) која се често користи код модела вештачких неуронских мрежа.

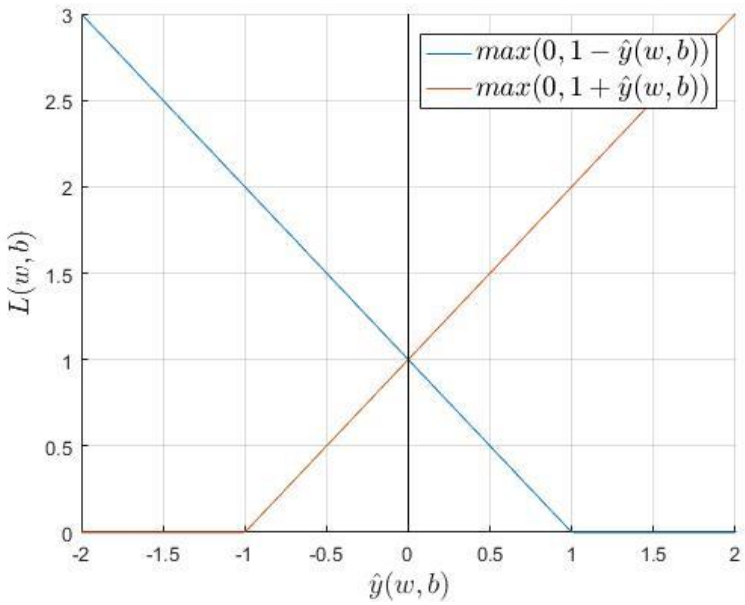
2.5.3.2.2.2.1 Функције грешке линеарне зависности

Функција грешке линеарне зависности се најчешће користи за обучавање класификационих модела базираних на максимизовању маргине класификације попут машине потпорних вектора. Ова функција грешке се може користити и код модела вештачких неуронских мрежа у спрези са функцијом активације хиперболичне тангенте вештачког неурона. Уколико посматрамо случај модела машине потпорних вектора, предвиђања модела $\hat{y}$ се могу формулисати као функција параметара w и b где се врши скаларни производ ортогоналног вектора хиперравни w и улазносг вектора x и додавање бијас параметра b на следећи начин:

$$\hat{y}(w,b) = w \cdot x + b$$

Циљ функције грешке линеарне зависности је максимизовање маргине класификације. Када су предвиђања модела $\hat{y}$ и тачне вредности лабела $y \in \{-1,1\}$ различитог знака, грешка предвиђања се линеарно повећава. У супротном, грешка тежи нули. Графичка репрезентација функције грешке линеарне зависности је приказана на Слика 25 а формула у наставку:

$$L(w,b) = \max_{\hat{y}}(0, 1 - y\hat{y}(w,b))$$



Слика 25 *Функција грешке линеарне зависности*

Уколико су тачна лабела y и предвиђена вредност $\hat{y}$ истог знака ($\max_{\hat{y}}(0, 1 - \hat{y}(w,b))$), грешка предикције $L(w,b)$ тежи нули што је предвиђена вредност $\hat{y}$ ближа

јединици. Што је предвиђена вредност $\hat{y}$ даља од вредности један онда се грешка предвиђања линеарно повећава. У другом случају када су тачна лабела y и предвиђена вредност $\hat{y}$ различитог знака знака ($\max_{\hat{y}}(0, 1 + \hat{y}(w, b))$), грешка предикције $L(w, b)$ се линеарно повећава што је предвиђена вредност $\hat{y}$ ближа јединици. Што је предвиђена вредност $\hat{y}$ ближа нули (али и даље супротног знака), онда се грешка смањује али је и даље велика (на примеру са Слика 25 у том случају је $L(w, b) \approx 2$).

Максимизовање маргине класификације у вишекласном задатку се може вршити на принципу више бинарних класификатора организованих по стратегији „један против једног“ или „један против свих“ (Секција 2.5.3.1.3). Уколико посматрамо случај стратегије „један против свих“, укупна грешка предвиђања се може рачунати као просечна грешка предвиђања по броју класа K на следећи начин:

$$L(w, b) = \frac{1}{K}\sum_{k=1}^{K} \max_{\hat{y}}(0, 1 - y_k \hat{y}_k(w, b))$$

Ради лакшег појашњења функције грешке унакрсне ентропије се за пример може користити математичка формулација модела једног вештачког неурона. Ако су улазни сигнали вештачког неурона означени са $x_1, x_2, \dots, x_n$, тежине синапси са $w_1, w_2, \dots, w_n$, бијас параметар са b, онда се излаз (предвиђања) вештачког неурона може дефинисати као функција на следећи начин:

$$\hat{y}(w, b) = f(w \cdot x + b)$$

Где $w \cdot x$ представља векторски производ улазних података x и тежина синапси w. Функција активације f у последњем слоју мреже се бира према типу класификације.

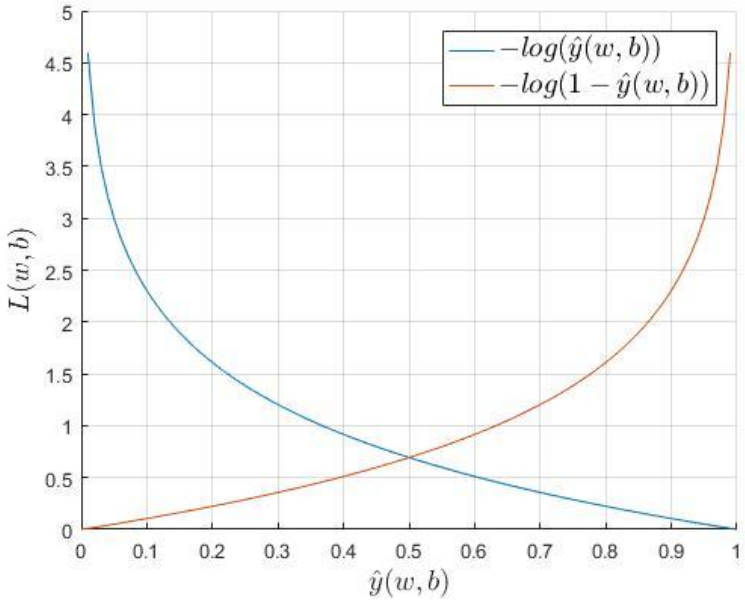
2.5.3.2.2.2.2 Функције грешке унакрсне ентропије

Функција грешке унакрсне ентропије подразумева одабир параметара w и b тако да се максимизује логаритам вероватноће тачне лабеле y. Уколико излаз вештачког неурона $\hat{y}$ посматрамо као функцију зависности параметара тежина синапси и бијаса $\hat{y}(w, b)$, онда се функција грешке унакрсне ентропије $L(w, b)$ у случају бинарне класификације дефинише на следећи начин:

$$L(w, b) = -(y \log \hat{y}(w, b) + (1 - y) \log(1 - \hat{y}(w, b)))$$

Уколико је тачна лабела y вредност јединице, онда је други део сабирка функције грешке $L(w, b)$ једнак нули и посматра се само део $-\log \hat{y}(w, b)$. Што је вредност

предвиђања $\hat{y}$ ближа јединици то грешка тежи нули а у супротном тежи бесконачности. У обратном случају, уколико је тачна лабела y вредност нуле, онда је први део сабирка функције грешке $L(w,b)$ једнак нули и посматра се само део $-\log(1-\hat{y}(w,b))$. Што је вредност предвиђања $\hat{y}$ ближа нули то грешка тежи нули а у супротном тежи бесконачности. На овај начин функција грешке максимизује грешке погрешних предвиђања и тиме указује моделу да су потребна нова знања како би модел био сигурнији. Графички приказ функције грешке унакрсне ентропије у оба случаја вредности тачне лабеле y је дат на Слика 26.



Слика 26 *Функција грешке унакрсне ентропије*

У случају вишекласне класификације се примена функције грешке унакрсне ентропије базира на суми појединачних грешака по класама. За разлику од бинарне класификације се на излазу модела вештачке неуронске мреже уместо једне вредности добија вектор предвиђања (вероватноћа). Такође, тачна лабела y представља вектор јединице где је вредност један на позицији одређене класе а остале вредности су нуле. Формално, уколико посматрамо проблем класификације K класа, тачне лабеле y_k и предвиђања модела $\hat{y}_k$ су вектори истих димензија K, онда се функција грешке унакрсне ентропије $L(w,b)$ дефинише на следећи начин:

$$L(w,b) = -\sum_{k=1}^{K} y_k \log \hat{y}_k(w,b)$$

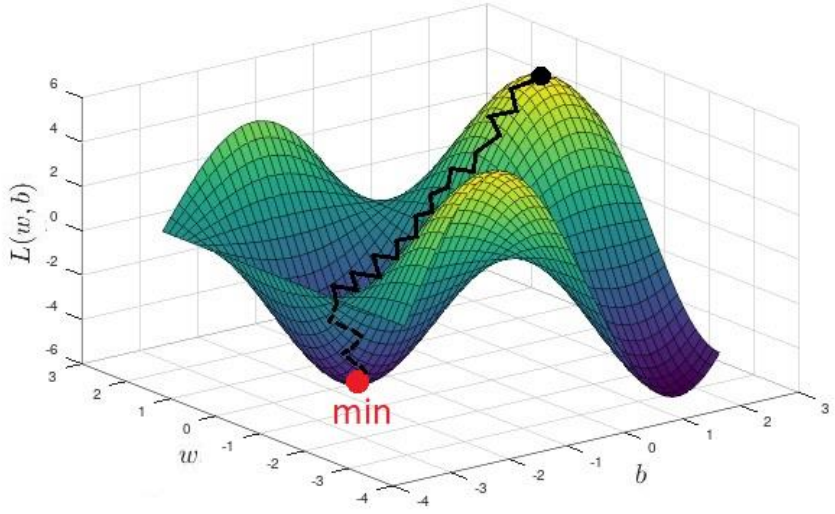
Са обзиром на чињеницу да је y_k вектор јединице који на k позицији има вредност један, укупна вредност фунцкије грешке ће садржати само грешку предвиђања за дату класу k. Сва остала предвиђања вектора $\hat{y}_k$ су анулирана нултим вредностима

вектора y_k. Што је вредност предвиђања $\hat{y}_k$ за дату класу k ближа јединици то укупна грешка тежи нули а у супротном тежи бесконачности (Слика 26).

2.5.3.2.2.3 Алгоритми оптимизације

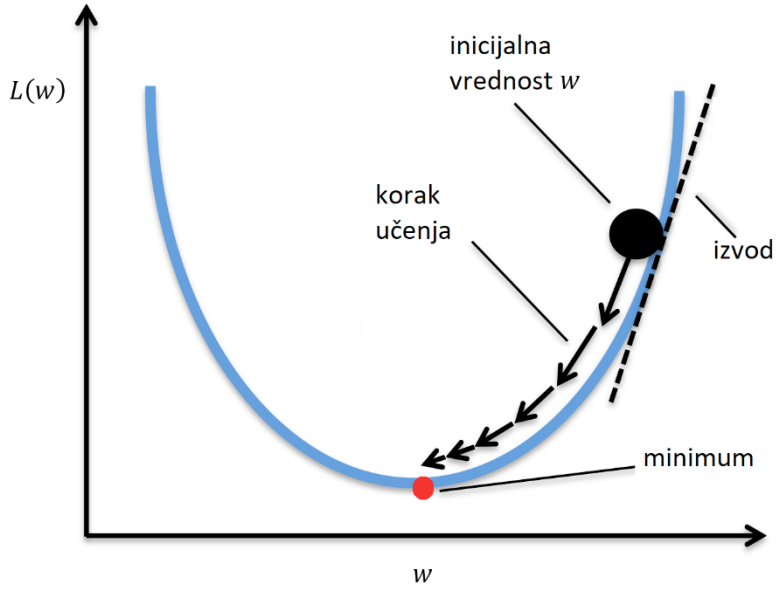
Алгоритми оптимизације служе за итеративну измену параметара модела тако да се минимизује функција грешке. Најпознатији алгоритам оптимизације је алгоритам опадајућег градијента (енгл. *Gradient Descent - GD*) који је објашњен у наставку. Поред *GD* алгоритма су описани и алгоритам опадајућег градијента са моментумом, прилагодљиви алгоритам градијента (енгл. *Adaptive Gradient Algorithm - Adagrad*), алгоритам пропагације корена квадратне средње вредности (енгл. *Root Mean Square Propagation - RMSProp*), алгоритам прилагодљиве естимације момента (егнл. *Adaptive Moment Estimation - Adam*), због уочених проблема брзине и стабилности конвергенције *GD* алгоритма.

Алгоритам опадајућег градијента је заснован на задатку минимизације функције грешке, односно на тражењу минимума дате функције. Тражење минимума функције се базира на рачунању градијента тј. парцијалног извода функције грешке у односу на сваки од параметара. У наставку ће бити узет пример модела једног вештачког неурона са једним улазом тј. два параметра w и b. Графички гледано, тражење минимума конвексне функције грешке се репрезентује у тражењу смера нагиба и померању за одређени корак у супротном смеру извода. Пример тражења минимума функције грешке $L(w, b)$ применом алгоритма опадајућег градијента је приказан на Слика 27.



Слика 27 *Пример алгоритма опадајућег градијента*

Тражење минимума у приказаном примеру (Слика 27) захтева измену параметара w и b у супротном смеру од вредности парцијалног извода функције грешке. Дати пример се може још поједноставити и свести на дводимензионални простор узимањем параметра b за константу. Пример алгоритма опадајућег градијента за функцију грешке са једним параметром $L(w)$ је приказан на Слика 28.



Слика 28 *Пример алгоритма опадајућег градијента за функцију са једним параметром*

Извод функције грешке из датог примера се рачуна као нагиб тангенте у почетној тачки тј. иницијалној вредности параметра w. Уколико је извод позитиван у датој тачки онда се за наредни корак опадајућег градијента узима трачка померена за одређни корак у супротном смеру извода тј. лево на графику (Слика 28). У другом случају, уколико је извод негативан у датој тачки онда се врши померање у десно на графику. Формално, измена вредности параметра $\hat{w}$ се врши на следећи начин:

$$\hat{w} = w - \alpha \frac{\partial}{\partial w} L(w)$$

Где параметар α представља корак учења (егнл. *learning rate*), односно магнитуду помераја опадајућег градијента. Вредностима параметра α контролишемо измену вредности параметра w. Већим вредностима параметра α дајемо већи утицај новијим информацијама (градијентима) у односу на стечене тј. старе инфромације које се чувају у параметру w. Из тог разлога параметар α метафорички представља брзину учења, ондосно брзину усвајања нових информација.

Када је парцијални извод функције грешке ($\frac{\partial}{\partial w}L(w)$) позитиван, онда се смањује вредност параметра $\widehat{w}$, а у супротном случају повећава. У наредном кораку алгоритма параметар w преузима измењену вредност $\widehat{w}$ и наставља се процес. С обзиром на то да конвергенција ка минимуму скоро никад није загарантована, процес се понавља до кад се не испуни следећи услов конвергенције:

$$\left|\frac{\partial}{\partial w}L(w)\right| \leq \epsilon$$

Алгоритам опадајућег градијента се зауставља када је апсолутна вредност градијента мања или једнака дефинисаној вредности ϵ. За вредност ϵ се уобичајно узима веома мала вредност, на пример 10^{-6}. То значи да се алгоритам зауставља уколико је напредак тражења минимума толико мали да је измена вредности параметра w занемарљива. У пракси се за услов конвергенције посматра и промена вредности грешке. Повећање вредности грешке указује на дивергенцију алгоритма и тада се зауставља процес тражења минимума функције грешке.

Уколико се постављени проблем зависности функције грешке једног параметра ($L(w)$) прошири на проблем зависности параметара w и b (Слика 27), онда се корак алгоритма опадајућег градијента састоји из тражења следећих парцијалних извода:

$$\widehat{w} = w - \alpha\frac{\partial}{\partial w}L(w,b)$$

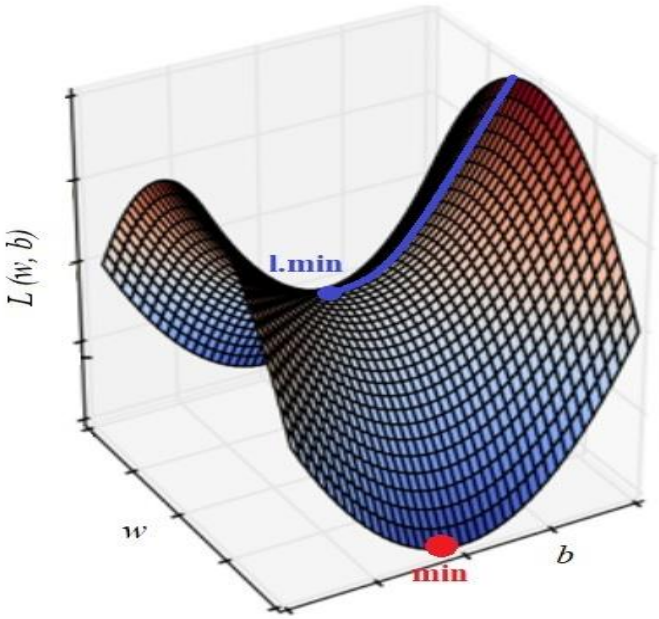
$$\hat{b} = b - \alpha\frac{\partial}{\partial b}L(w,b)$$

У реалнијем сценарију алгоритма опадајућег градијента са N параметара у једном слоју вештачке неуронске мреже, променљиве w и b представљају векторе N димензија за које се рачунају парцијални изводи вектора $\widehat{w}$ и $\hat{b}$ истих димензија. Повећањем броја скривених слојева се повећава број параметара које је потребно изменити („научити“) у што краћем времену.

Време обучавања модела тј. брзина оптимизационог алгоритма зависи од више фактора као што су корак учења и величина обучавајућег скуп података. Приликом одабира корака учења је потребно обратити пажњу јер у случају превише великих вредности може доћи до дивергенције тј. кретању оптимизационог алгоритма даље од минимума. У супротном случају, када је корак учења сувише мала вредност, онда се успорава процес обучавања и потребно је знатно више времена за обучавање. Стога је и вредност корака учења потребно оптимизовати.

На време обучавања модела такође утиче и величина обучавајућег скупа података која се користи за рачунање једног корака оптимизационог алгоритма. Када се

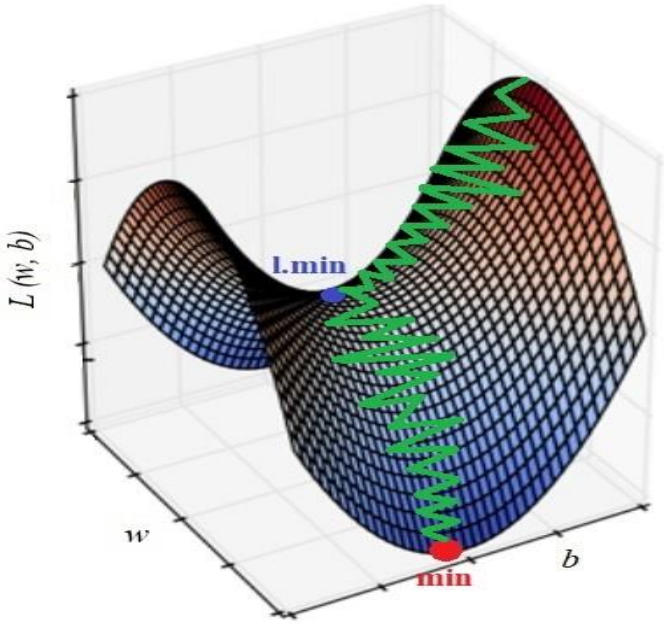
за једну епоху учења, односно један корак *GD* оптимизационог алгоритма, користи цео обучавајући скуп података (енгл. *batch*) онда то доводи до линеарне зависности времена обучавања и величине обучавајућег скуп података. То значи да се кроз модел пропусти цео обучавајући скуп података, израчуна просечна вредност градијента, и након тога изврши измена параметара модела w. Предност коришћења целог обучавајућег скупа података јесте стабилна конвергенција до оптималних вредности параметара w. Такође, рачунски ефикасније јер се измена парамтера врши само једном за целу епоху. Мана овог приступа је спора конвергенција оптимизационог алгоритма због линеарне зависности времена обучавања и величине обучавајућег скупа података. Уколико функција грешке има више минимума, *GD* оптимизациони алгоритам услед стабилне конвергенције и споре измене параметара w не успева да заобиђе локални минимум и пронађе глобални (Слика 29).



Слика 29 *Проблем локалног минимума за GD оптимизациони алгоритам*

Обучавајући скуп података се може поделити на подскупове (енгл. *mini-batch*) чиме ће процес учења тј. измене вредности параметара модела бити учесталији. Уколико је величина подскупа најмања тј. садржи само један обучавајући податак, онда се тај приступ назива и стохастички приступ *GD* оптимизационог алгоритма (енгл. *Stochastic GD - SGD*). У *SGD* приступу се након сваког обучавајућег податка рачуна грешка предикције и врши измена вредности параметара модела. Предност овог приступа је бржа конвергенција оптимизационог алгоритма тј. брже усвајање (учење) нових информација из података. Међутим, приступи са мањим подскупом обучавајућих података нису отпорни на шум у подацима и доводе до тога да кораци оптимизационог

алгоритма не буду увек у смеру минимума. Тачније, вредност грешке предикције се неће увек смањивати током учења него ће варирати (осцилирати). Ова мана може представљати и предност *SGD* оптимизационог алгоритма у случају постојања више минимума. Осцилирањем *SGD* оптимизационог алгоритма приликом конвергенције и учесталијом променом параметара w се омогућава мимоилажење локалних минимума и проналазак глобалног минимума (Слика 30).



Слика 30 *SGD оптимизационим алгоритмом*

Смањењем величине подскупова, односно повећањем броја подскупова се повећава фреквенција измене вредности параметара w што представља рачунски захтевне операције. Стога је за процес учења, поред корака учења, веома битно прилагодити величину подскупова како би се обезбедила већа отпорност оптимизационог алгоритма на шум у подацима а са друге стране бржа конвергенција.

Једно од решења за бржу конвергенцију *SGD* оптимизационог алгоритма на мањим подскуповима обучавајућих података је примена *SGD* алгоритма са моментумом. Овај алгоритам се базира на статистичком прорачуну покретног просека података са експоненцијалним утицајем додатних тежина (егнл. *exponentially weighted average*). Приликом измене вредности параметара синаптичких тежина w и бијаса b се врши множење корака учења са тежинским фактором који се експоненцијално смањује током итерација алгоритма. Ако је функција грешке дефинисана као $L(w, b)$, измена параметара w и b се врши на основу следећих формула:

$$\hat{v}_w = \beta v_w + (1-\beta)\frac{\partial}{\partial w}L(w,b)$$

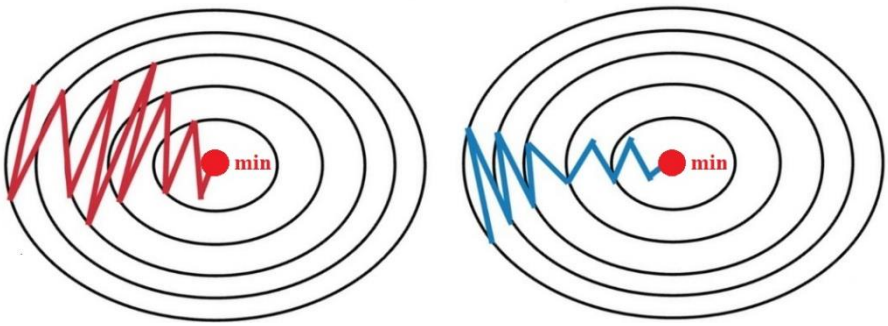
$$\hat{w} = w - \alpha\hat{v}_w$$

$$\hat{v}_b = \beta v_b + (1-\beta)\frac{\partial}{\partial b}L(w,b)$$

$$\hat{b} = b - \alpha\hat{v}_b$$

На основу параметра β се врши контрола утицаја одређеног броја претходних измена тј. працијалних извода v на тежински просек $\hat{v}$. Већим вредностима параметра β се повећава утицај суме претходних парцијалних извода v на тежински просек $\hat{v}$ а смањује утицај тренутног градијента. Типичне вредности параметра β су веома близу вредности један (нпр. 0,9 или 0,99). Измене параметара w и b се врше у суротном смеру од тежинског просека $\hat{v}_w$ и $\hat{v}_b$, помножених са кораком учења α. У наредној итерацији, вредности параметара $\hat{v}_w$ и $\hat{v}_b$ се смештају у параметре v_w и v_b и понављају исти кораци.

На овај начин се подацима који представљају шум не дозвољава да пуно утичу на измену параметара w и b уколико је промена знатно виша од просека претходних измена. Графички гледано, *SGD* алгоритам са моментумом (Слика 31, десно) знатно брже конвергира у односу на базни *SGD* алгоритма (Слика 31, лево) јер се параметром β смањује моменат силе (моментум) дотадашњег кретања алгоритма и повећава утицај новијих парцијалних извода. Тиме се омогућава превазилажење превојних тачака функције грешке у којима је сума претходних парцијалних извода v веома мала.



Слика 31 *SGD са моментумом*

Mana *SGD* алгоритма са моментумом је то што се измена параметара w не врши равномерно, што доводи до већих осцилација у једном смеру конвергенције а мањих у другом. Решење овог проблема је представљено *Adagrad* алгоритмом где се вредности градијената своде на сличне распоне вредности. Измена параметара w и b се базира на дељењу градијената кореном суме квадрираних градијената. Формалније, ако се прарцијални изводи функције грешке $L(w,b)$ по параметрима w и b означе са ∇w и ∇b,

суме квадрираних парцијалних извода са $\hat{v}_w$ и $\hat{v}_b$, онда се измена параметара w и b врши на следећи начин:

$$\nabla w = \frac{\partial}{\partial w} L(w,b); \quad \nabla b = \frac{\partial}{\partial b} L(w,b)$$

$$\hat{v}_w = v_w + (\nabla w)^2$$

$$\hat{w} = w - \alpha \frac{\nabla w}{\sqrt{\hat{v}_w} + \varepsilon}$$

$$\hat{v}_b = v_b + (\nabla b)^2$$

$$\hat{b} = b - \alpha \frac{\nabla b}{\sqrt{\hat{v}_b} + \varepsilon}$$

Квадрирањем парцијалних извода ∇w и ∇b је повећан утицај новијих градијената на коначне суме градијената $\hat{v}_w$ и $\hat{v}_b$. Измена парматера w и b се врши на основу парцијалних извода ∇w и ∇b подељених кореном суме квадрираних градијената $\hat{v}_w$ и $\hat{v}_b$, помножених са кораком учења α. Како би се пречило дељење нулом, на вредност корена сума квадрираних градијената $\hat{v}_w$ и $\hat{v}_b$ се додаје параметар ε чија вредност је веома мала (реда 10^{-8}).

Велике вредности градијената се деле великим вредностима корена, чиме се смањује утицај последњих градијената на измене параметара $\hat{w}$ и $\hat{b}$. У другом случају, мале вредности градијената остају и даље мале чиме се даје допринос измени параметра $\hat{w}$ и $\hat{b}$. На овај начин је измена вредности параметра $\hat{w}$ и $\hat{b}$ равномернија што убрзава алгоритам оптимизације у односу на *SGD* алгоритам са моментумом.

Међутим, дељењем градијената превеликим вредностима у *Adagrad* алгоритму се успорава измена вредности параметра $\hat{w}$ и $\hat{b}$ у дубоким вештаким неуронским мрежама. Решење овог проблема *Adagrad* оптимизационог алгоритма је представљено у *RMSprop* алгоритму са додатим моментумом. Формално, ако се прарцијални изводи функције грешке $L(w,b)$ по параметрима w и b означе са ∇w и ∇b, суме квадрираних парцијалних извода са моментумом као $\hat{v}_w$ и $\hat{v}_b$, онда се измена параметара w и b врши на начин дефинисан у *Adagrad* алгоритму:

$$\nabla w = \frac{\partial}{\partial w} L(w,b); \quad \nabla b = \frac{\partial}{\partial b} L(w,b)$$

$$\hat{v}_w = \beta v_w + (1-\beta)(\nabla w)^2$$

$$\hat{w} = w - \alpha \frac{\nabla w}{\sqrt{\hat{v}_w} + \varepsilon}$$

$$\hat{v}_b = \beta v_b + (1-\beta)(\nabla b)^2$$

$$\hat{b} = b - \alpha \frac{\nabla b}{\sqrt{\hat{v}_b} + \varepsilon}$$

Квадрирањем парцијалних извода се ублажава смањење градијента дељењем великим бројем, те не долази до успоравања измене параметара $\hat{w}$ и $\hat{b}$.

Оптимизациони алгоритам који комбинује предности претходно описаног *SGD* алгоритма са моментумом и *RMSprop* алгоритма је *Adam* алгоритам. Измене параметара $\hat{w}$ и $\hat{b}$ се врше на приступу описаном у *RMSprop* алгоритму где су имениоци тј. парцијални изводи ∇w и ∇b замењени сумама градијената $\hat{v}_{w_m}$ и $\hat{v}_{b_m}$ из *SGD* алгоритма. Формуле по којој се рачунају промене параметара $\hat{w}$ и $\hat{b}$ су дате у наставку:

$$\nabla w = \frac{\partial}{\partial w} L(w,b); \quad \nabla b = \frac{\partial}{\partial b} L(w,b)$$

$$\hat{v}_{w_m} = \beta_1 v_{w_m} + (1-\beta_1)\nabla w$$

$$\hat{v}_{b_m} = \beta_1 v_{b_m} + (1-\beta_1)\nabla b$$

$$\hat{v}_{w_{rms}} = \beta_2 v_{w_{rms}} + (1-\beta_2)(\nabla w)^2$$

$$\hat{v}_{b_{rms}} = \beta_2 v_{b_{rms}} + (1-\beta_2)(\nabla b)^2$$

$$\hat{w} = w - \alpha \frac{\hat{v}_{w_m}}{\sqrt{\hat{v}_{w_{rms}}} + \varepsilon}$$

$$\hat{b} = b - \alpha \frac{\hat{v}_{b_m}}{\sqrt{\hat{v}_{b_{rms}}} + \varepsilon}$$

На основу параметара β_1 и β_2 се врши контрола утицаја одређеног броја претходних измена тј. працијалних извода v на тежински просек $\hat{v}$. Типичне вредности параметара β_1 и β_2 су веома близу вредности један (нпр. 0,9 и 0,99). На овај начин се обезбедило убрзање конвергенције у односу на *Adagrad* алгоритам. Међутим, на почетку рада *Adam* алгоритма је уочена спора измена параметара $\hat{w}$ и $\hat{b}$ због малих вредности градијената који се додају на тежинске просеке $\hat{v}$. Малим вредностима $(1-\beta_1)$ и $(1-\beta_2)$ се смањује вредност нових градијената, чиме се успорава измена тежинских сума $\hat{v}$. Из тог разлога је иницијални предлог *Adam* алгоритма допуњен изменама претходно израчунатих тежинских просека $\hat{v}$ на следећи начин:

$$\hat{v}_{w_m} = \frac{\hat{v}_{w_m}}{(1-\beta_1^t)}$$

$$\hat{v}_{b_m} = \frac{\hat{v}_{b_m}}{(1-\beta_1^t)}$$

$$\hat{v}_{w_{rms}} = \frac{\hat{v}_{w_{rms}}}{(1-\beta_2^t)}$$

$$\hat{v}_{b_{rms}} = \frac{\hat{v}_{b_{rms}}}{(1-\beta_2^t)}$$

Тежински просеци $\hat{v}$ се деле разликама $(1-\beta_1^t)$ и $(1-\beta_2^t)$ које зависе од параметра итерације алгоритма t. На почетку рада алгоритма делиоци тежинских просека $\hat{v}$ су велики и на тај начин анулирају утицај на нове градијенте. Дакле у првој итерацији алгоритма, када параметар t има вредност један, тежински просеци $\hat{v}$ су измењени на следећи начин:

$$\hat{v}_{w_m} = \beta_1 v_{w_m} + \nabla w$$

$$\hat{v}_{b_m} = \beta_1 v_{b_m} + \nabla b$$

$$\hat{v}_{w_{rms}} = \beta_2 v_{w_{rms}} + (\nabla w)^2$$

$$\hat{v}_{b_{rms}} = \beta_2 v_{b_{rms}} + (\nabla b)^2$$

У наставку рада алгоритма, када се вредности параметра t повећавају, тада се смањује утицај делиоца тежинских просека $\hat{v}$ и *Adam* алгоритам ради по иницијалном предлогу. Резултат ове измене *Adam* алгоритма је бржа конвергенција током свих итерација у односу на претходно описане оптимизационе алгоритме. Међутим, у завиности од проблема, оптимизациони алгоритми имају различите карактеристике те је стога потребно извршити проверу и других оптимизационих алгоритама за конкретан проблем.

2.5.3.2.2.3.1 Оптимизација хипер-параметара

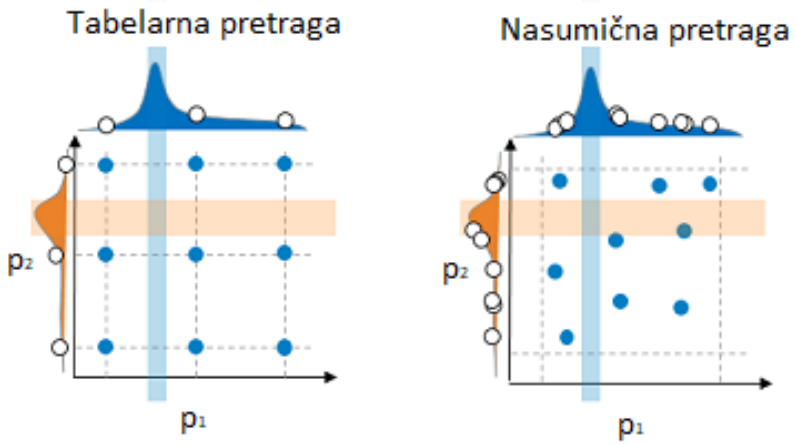
Хипер-параметри (енгл. *hyper-parameters*) модела се користе за контролу процеса учења параметара модела машинског учења попут тежина и бијаса вештачке неуронске мреже. Циљ учења је пронаћи комбинацију вредности хипер-параметара на основу којих модел машинског учења успева да научи оптималне вредности параметара модела за решавање одређеног проблема. Дакле, током учења је потребно оптимизовати вредности хипер-параметара тако да модел постиже најбоље перформансе на конкретном задатку. Оптимизација хипер-параметара се врши на

обучавајућем скупу података, а евалуација на валидационом скупу података (Секција 2.4).

Хипер-параметри модела машинског учења засвисе од конкретног модела. На пример, за модел машине потпорних вектора (Секција 2.5.3.1.3) се у хипер-параметре убрајају одабир функције језгра, параметар регулације, специфични парамтери за одређену функцију језгра, и тако даље. У хипер-параметре модела вештачке неуронске мреже се убарају број скривених слојева, број вештачких неурона у слоју, избора активационе функције, избора функције грешке, избора оптимизационог алгоритма, избора корака учења оптимизационог алгоритма, избора величине подскупа обучавајућих података, избора типа регуларизације (Секција 2.5.3.2.2.4.1), и тако даље.

Оптимизација хипер-параметара се генерално не врши аутоматски за све хипер-параметре. На пример, тип активационе функције, функције грешке, архитектура неуронске мреже, или вредности хипер-параметара β_1, β_2 и ε код *Adam* оптимизационог алгоритма се у већини случајева постављају ручно на основу претходних искустава пронађених у литератури. Док се за хипер-параметре попут корака учења, величине подскупа обучавајућег скупа и параметра регуларизације користе посебни алгоритми за аутоматску оптимизацију.

Алгоритми оптимизације хипер-параметара се према начину тражења оптималних вредности хипер-параметара деле на приступе табеларне претраге (енгл. *grid search*), насумичне претраге (енгл. *random search*) и примени алгоритма машинског учења. Приступ претраге табеле користи скуп могућих комбинација, унапред предефинисаних, вредности параметара (Слика 32, лево). За сваки хипер-параметар се дефинише скуп могућих вредности. Затим се обучава модел на обучавајућем скупу за сваку комбинацију вредности хипер-параметара и врши евалуација перформанси на валидационом скупу података. Коначна комбинација вредности хипеп-параметара је она за коју модел остварује најбоље перформансе на валидационом скупу. Уколико се дефинише довољно велик скуп могућих вредности хипер-параметара, овај приступ ће сигурно пронаћи најбољу конбинацију. Међутим, овај приступ је рачунски захтеван јер време извршавања експоненцијално расте повећањем броја хипер-парметара и броја могућих вредности хипер-параметара. На пример, ако обучавање модела са два хипер-параметара траје један сат, онда време које је потребно за табеларну претрагу три могуће вредности за сваки хипер-параметар износи девет сати (3^2).



Слика 32 *Табеларна и насумична претрага вредности параметара*[14]

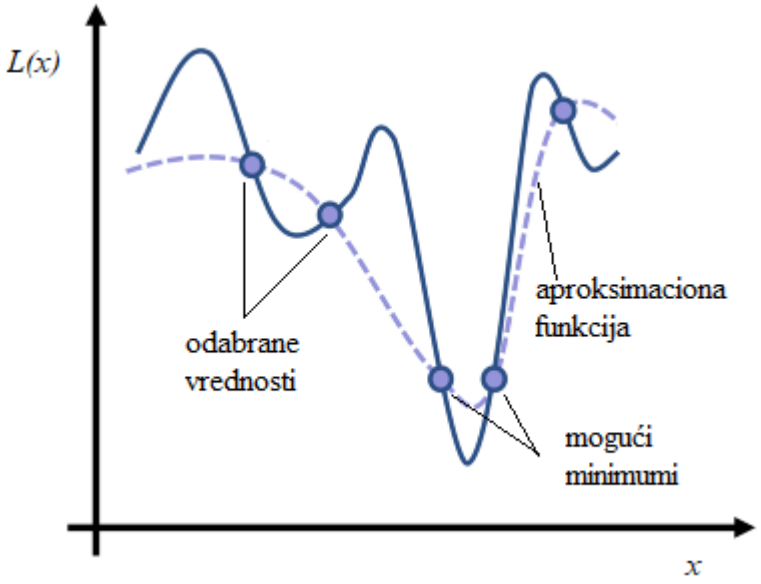
Приступ насумичне претраге врши насумично генерисање вредности за одређене параметре на основу дефинисане функције и опсега (Слика 32, десно). За дефинисан број могућих вредности хипер-параметара се насумично бирају вредности из одређеног опсега вредности добијених на основу одређене расподеле (нпр. нормална, биномна). На крају дефинисаног максималног броја итерација примене насумичног одабира вредности хипер-параметара на обучавајућем скупу се добија комбинација хипер-параметара за коју модел остварује најбоље перформансе на валидационом скупу. Дефинисањем већег броја итерација се повећава вероватноћа да се пронађе најбоља комбинација хипер-параметара.

Међутим, због насумичног избора вредности овај приступ не гарантује проналазак оптималних вредности хипер-параметара ни после великог броја итерација. Такође, на насумичан избор комбинације хипер-параметара не утичу перформансе модела за претходне изборе хипер-вредности. Из тог разлога се веома често примењују приступи машинског учења попут Бајесове оптимизације (енгл. *Bayesian Optimization*).

Бајесова оптимизација посматра проблем претраге вредности хипер-параметара као проблем оптимизације перформанси модела. Одабир вредности хипер-параметара зависи од претходног искуства тј. вредности оптимизационе функције. На почетку се насумично одаберу вредности хипер-параметара, обучи модел на обучавајућем скупу и изврши евалуација на валидационом скупу. Вредности за коју модел даје најбоље перформансе представљају могуће минимуме оптимизационе функције. У наредном кораку се даље насумично бирају вредности из области вредности могућих минимума и врши кориговање модела.

[14] Извор слике: https://polyaxon.com/

Учењем модела се формира функција која апроксимира оптимизациону функцију, познатију и као сурогат функција. Сурогат функција је репрезентована Гаусовим процесом где се уместо вредности чувају вероватноће које означавају колико побољшања перформанси модела дониси одређена комбинација вредности хипер-параметара. Пример рада Бајесове оптимизације једног параметра је дата на Слика 33.



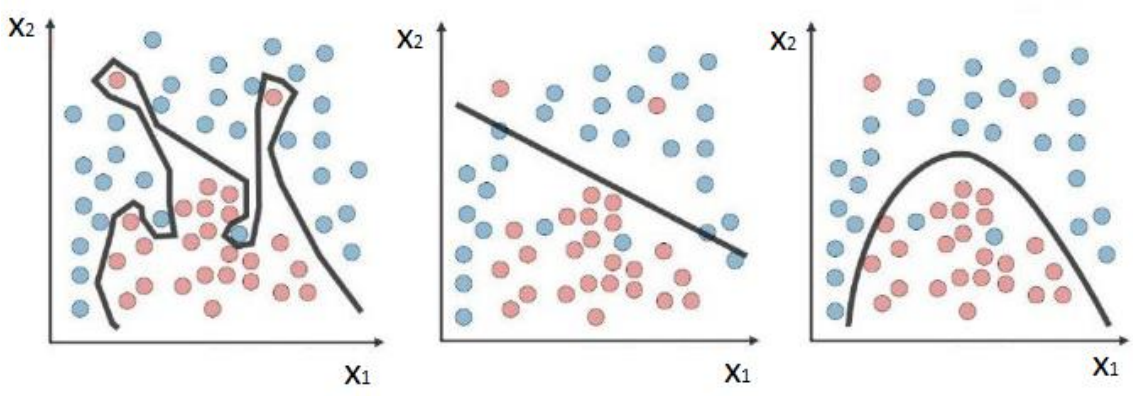
Слика 33 *Бајесова оптимизација*

У овом истраживању су коришћени сва три приступа оптимизације хипер-параметара модел машинског учења.

2.5.3.2.2.4 Проблем преобучавања и подобучавања

Приликом обучавања модела машинског учења постоји проблем преобучавања (енгл. *overfitting*) и подобучавања (енгл. *underfitting*). Проблем преобучавања подразумева сувише добро прилагођавање модела обучавајућим подацима чиме се губи могућност генерализације података. То доводи до добрих перформанси модела на обучавајућем скупу података али лошијих перформанси на тест скупу података. Разлог је у томе што се модел сувише добро прилагодио подацима из обучавајућег скупа који представљају само узорак свих могућих података датог проблема. У обратном случају, када се модел недовољно прилагодио подацима из обучавајућег скупа, представља проблем подобучавања. Тада је модел лоше обучен и остварује слабе перформансе на обучавајућем скупу података.

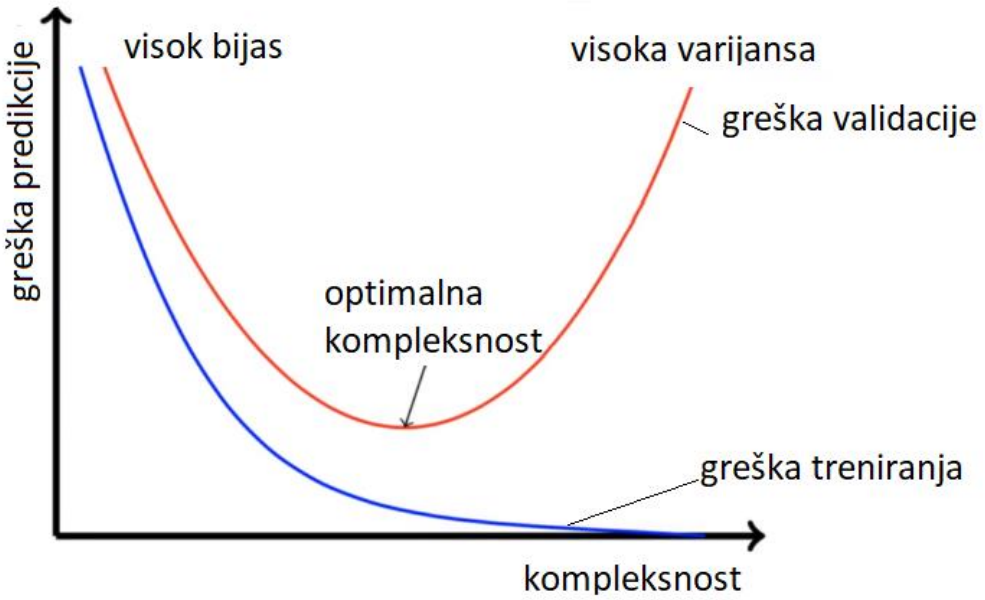
Решење преобучавања и подобучавања је у средини, односно у обучавању модела да довољно добро генерализује обучавајуће податке како би остварио боље перформансе над подацима које није сусрео током обучавања. На Слика 34 је приказан пример преобучавања, подобучавања и модела који је довољно добро обучен за пример бинарне класификације.



Слика 34 *Пример преобучавања, подобучавања и оптималног обучавања модела*

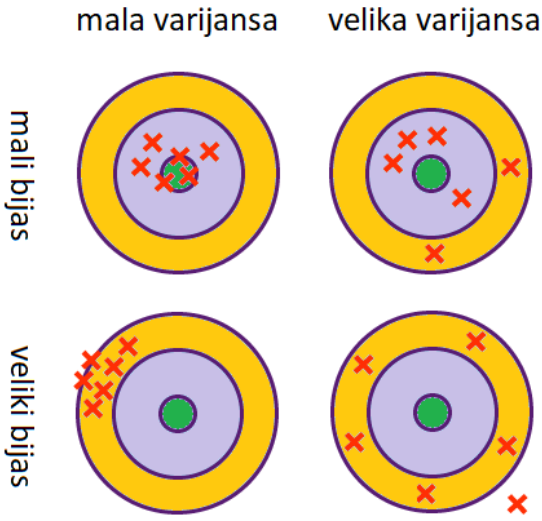
На примеру са Слика 34 се може видети да је хиперраван раздвајања дводимензионалног простора две класе у случају преобучавања сувише добро прилагођена подацима. Насупрот томе, хиперраван подобученог модела не раздваја довољно добро простор податке две класе. У крајњем примеру је се може видети оптимална хиперраван која довољно добро раздваја простор података две класе.

Проблем преобучавања и подобучавања модела се често пресликава на однос тј. компромис мере бијаса и варијансе грешке предвиђања модела у односу на комплексност модела (Слика 35). Што се модел дуже обучава то се повећава комплексност хиперравни модела која раздваја простор обучавајућих података на класе. Бијас је мера колико су предвиђања далеко од исправних вредности генерално током обучавања и валидације модела. Када модел има велику грешку предвиђања и на обучавајућем и валидационом скупу података то значи да је модел подобучен и има велику вредност бијаса. Варијанса је мера доследности, односно варијабилности, модела приликом предвиђања одређеног узорка податка. Уколико је варијанса модела велика то занчи да модел није сигуран и да је осетљив на случајност у подацима. У овом случају модел је преобучен и тешко тачно предвиђа нове податке. Оптимална комплесност модела се постиже када је однос бијаса и варијансе усклађен тј. када су њихове вредности мале.



Слика 35 *Однос бијаса и варијансе модела*

На илустративном примеру обучавања модела за погађање у мету се може видети однос мере бијаса и варијансе модела (Слика 36). Може се закључити да није довољно да модел буде само поуздан (мала варијанса) него је за погађање централног дела мете потребна и прецизност (мали бијас).



Слика 36 *Илустративан пример односа бијаса и варијансе модела*

Проблем преобучавања модела се може уочити по великој варијанси, малој грешци предвиђања приликом учења која је знатно мања од грешке предвиђања на валидационом скупу података. Преобучавање модела се може решити применом једног или више следеће наведених начина:

- регуларизацијом параметра модела, која је објашњена у наставку
- повећањем обучавајућег скуп података чиме би повећао узорак могућих података датог проблема
- променом архитектуре модела

Модел који је подобучен има велики бијас и велику грешку на тернинг и валидационом скупу података. Подобучавање модела се може решити применом једног или више следеће наведених начина:

- додавањем више карактеристика чиме би се боље уочиле разлике између класа
- повећањем комплексности модела (нпр. повећањем броја скривених слојева у вештачким неуронским мрежама)
- дужим обучавањем модела чиме би се покушало боље прилагођавање модела теренинг скупу података

Код оптималног модела варијанса и бијас су усклађени, односно грешка предвиђања током обучавања је нешто нижа од грешке предвиђања током валидирања модела.

2.5.3.2.2.4.1 Регуларизација

Регуларизација представља метод решавања проблема преобучавања односно сувише доброг прилагођавања модела обучавајућим подацима. Постоји више начина примене регуларизације. Један од начина је примена регуларизације у оквиру функције грешке.

Циљ регуларизације је да смањи укупну грешку модела на обучавајућем скупу и да смањи велике вредности параметара модела. Регуларизацијом се врши кажњавање великих вредности параметра повећањем грешке предвиђања чиме се спречава случај преобучавања модела. У оквиру секције модела машине потпорних вектора је описана функција грешке линеарне зависности са регуларизационим делом једначине (Секција 2.5.3.1.3). Формула функција грешке је приказана на следећи начин:

$$L = \frac{1}{n}\sum_{i=1}^{n} max(0, 1 - y_i \cdot f(x_i;\, w_i)) + \gamma \frac{1}{2n}\sum_{i=1}^{n} w_i^2$$

Регуларизациони део једначине без скалирајућих константи се може представити као функција $R(w)$ на следећи начин:

$$R(w) = \gamma \sum_{i=1}^{n} w_i^2$$

Приказана функција регуларизације представља другу Еуклидову норму (Л2 норма) која вредности тежина w скалира по квадратној функцији. Поред Л2 номре се може користити и прва Менхетн норма (Л1 норма) која вредности тежина w скалира по апсолутној функцији. Функција Л1 регуларизације је приказана у наставку:

$$R(w) = \gamma \sum_{i=1}^{n} |w_i|$$

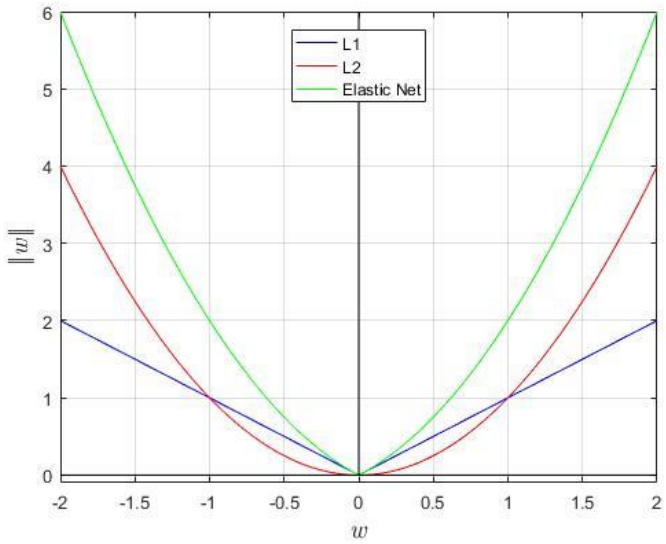
Параметар γ представља хипер-параметар модела којим се одређује утицај једначине регуларизације $R(w)$ на укупну функцију грешке L. Типичне вредности γ параметра су у опсегу $[0, \infty]$. Великим вредностима параметра γ се смањују вредности параметра w, где за много велике вредности γ долази до отежаног учења модела тј. подобучавања. Код неуронских мрежа велике вредности γ параметра доводе до тога да се одређени вештачки неурони понашају линеарно тј. постају пасивнији и спорије врше додатна учења (допуне тежина w). Мањим вредностима γ параметра се смањује утицај регуларизације на функцију грешке што може довести до преобучавања модела. Стога се хипер-параметар γ оптимизује посебно за одређени задатак (Секција 2.5.3.2.2.3.1).

Разлике Л1 и Л2 регуларизације се односе на начин како утичу на вредности тежина w. Л1 регуларизација тежи смањењу вредности тежина w на нулу, док се Л2 регуларизација тежи једнаком смањењу вредности тежина w. Другим речима, Л1 регуларизација тежи постављању медијана у подацима док Л2 регуларизација тежи постављању просека података. Стога је Л1 регуларизација корисна за селекцију великог броја карактеристика где се елиминишу оне карактеристике које нису битне тј. чије вредности су нуле. Док је Л2 регуларизација корисна када постоји колинеарност карактеристика тј. када је већина карактеристика битна. Комбинацијом предности Л1 и Л2 норме се добија еластична мрежа (енгл. *Elastic Net*) која линеарно комбинује утицај Л1 и Л2 регуларизације. Функција регуларизације еластичне мреже је приказана у наставку:

$$R(w) = \gamma \sum_{i=1}^{n} w_i^2 + (1 - \gamma) \sum_{i=1}^{n} |w_i|$$

Вредности параметра γ су у опсегу $[0, 1]$ чиме се дефинише утицај Л1 и Л2 регуларизације на функцију грешке. Са граничним вредностима опсега γ параметра се

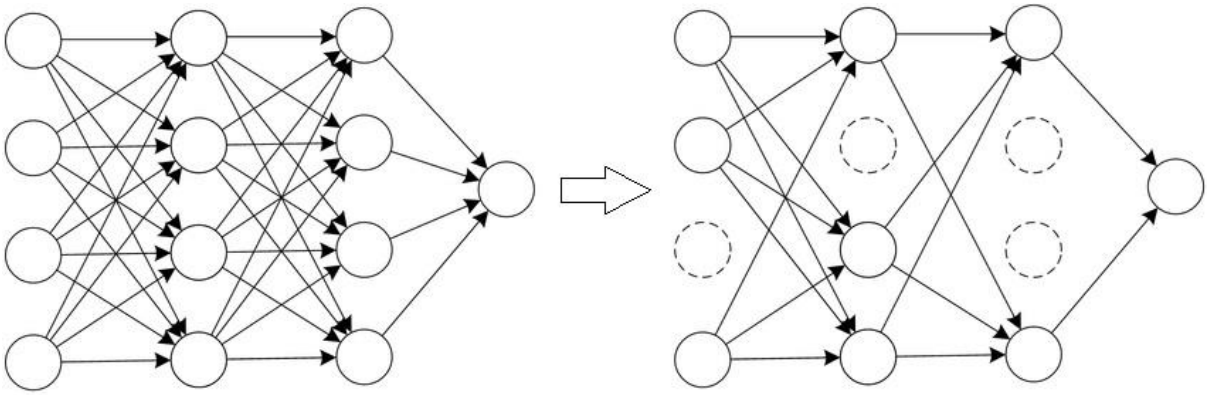
искључује утицај одређене регуларизације. Када је вредност γ параметра нула, онда се добија Л1 регуларизација. У супротном, на функцију грешке се додаје Л2 регуларизација. Графичка репрезентација утицаја типова регуларизација на вредности параметра w је дата на Слика 37.



Слика 37 *Типови регуларизације функције грешке*

Други начин реуларизације вештачких неуронских мрежа подразумева деактивацију вештачких неурона на основу одређеног прага вероватноће из опсега $(0, 1)$ (енгл. *dropout*). Услед великог броја вештачких неурона модел има велику флексибилност да се сувише прилагоди обучавајућим подацима (случај преобучавања). Циљ ове врсте регуларизације јесте спречавање преобучавања тиме што се мрежа обучава на тај начин да се не ослања на све вештачке неуроне мреже а да постигне добре резултате.

У оквиру хипер-параметара модела вештачке неуронске мреже се дефинише горњи праг вероватноће за деактивацију вештачких неурона. Затим се током обучавања модела врши насумично деактивирање вештачких неурона чија излазна вредност (вероватноћа) је испод дефинисаног прага. Технички гледано, током једне итерације учења се за сваки слој вештачке неуронске мреже генеришу насумичне вредности из опсега $[0, 1]$. Дате вредности се затим пореде са дефинисаним прагом вероватноће и тиме формира маска нула и јединица. Затим се излази неурона множе са датом маском чиме се добија да су излази неких неуорна нуле тј. неурон је неактиван и не утиче на учење модела у датој итерацији. У наредној итерацији се понавља процес и деактивирају други вештачки неурони случајним избором. Резултат деактивације вештачких неурона доводи до измене архитектуре модела (Слика 38).



Слика 38 *Деактивација вештачких неурона на основу прага вероватноће*

Међутим, модел обучаван на овакав начин није могуће валидирати на исти начин као и типичне вештачке неуронске мреже. Деактивирањем неурона модел није обучен да ради са свим неуронима тј. великим вредностима улаза неурона. Стога је у случају валидације модела потребно скалирати активације вештаких неурона за дефинисан праг вероватноће како би се на излазу добили очекивани излази као током обучавања.

Још један од начина регуларизације подразумева прекидање процеса обучавања у оптималном тренутку комплексности модела (Слика 35). Овај начин се базира на праћењу вредности грешке на обучавајућем и валидационом скупу. Што се дуже модел обучава, грешка на обучавајућем скупу опада до вредности близу нуле, што указује на случај преобучавања. Насупрот томе, грешка модела на валидационом скупу прво почиње да пада и након неког тренутка почиње да расте. Разлог тога је што преобучен модел није довољно флексибилан за примере које није сусрео током обучавања. Циљ је зауставити обучавање модела када грешка модела на валидационом скупу почне да расте. Технички гледано, током обучавања се након одређеног броја епоха сачувају копије модела које се касније валидирају. Оптимални модел је онај који има најмању грешку на обучавајућем и валидационом скупу података.

2.5.3.2.3 Дубоке вештачке неуронске мреже

Архитектуре вештачких неуронских мрежа које имају пуно слојева са пуно вештачких неурона се називају дубоким вештачким неуронским мрежама (енгл. *deep artificial neural networks*) или краће дубоким учењем. Дубоке вештачке неуронске мреже представљају моћну машину за учење која одговара *NLP* задацима. Главна компонента вештачких неуронских мрежа за решавање *NLP* задатака је употреба слоја за репрезентацију речи као векторе реалних вредности (енгл. *embedding layer*). У овом

слоју се врши маприање речи, као скуп дискретних симбола, на векторе вредности у релативно малом векторском простору са којим машина може радити. Удаљеност између вектора у векторском простору се може изједначити са растојањем између речи у језичком смислу (простору), што олакшава генерализацију понашања из једног у други простор (Goldberg, 2017). На основу векторске репрезентације речи вештачка неуронска мреже током процеса обучавања учи да комбинује векторе речи на начин који је користан за предвиђање. На овај начин се до неке мере ублажава проблем недостатка и дискретности података. Пример поређења вектора на основу њихове аналогије је описан у Секцији 2.2.1.3.1 на примеру Слика 11.

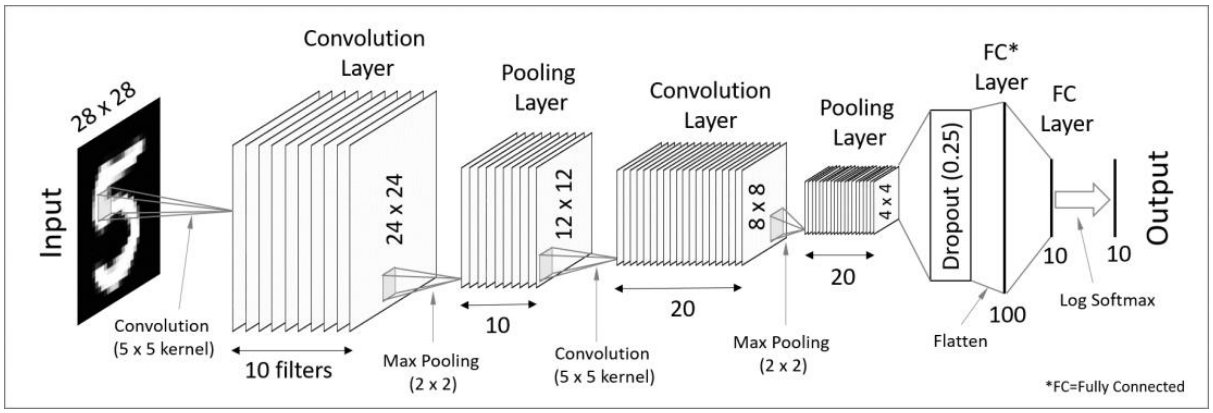
Вештачке неуронске мреже архитектуре вишеслојног перцептрона (Секција 2.5.3.2.1) омогућава рад са улазима фиксне дужине. Ови типови архитектуре вештачких неуронских мрежа се нису добро показале на одређеним *NLP* задацима. Проблем је немогућност моделовања дугачких зависности тј. уочавање дужег контекста и редоследа речи. За моделовање дугачких зависности је потребно доста више вештачких неурона што повећава комплексност тј. број параметара које је потребно оптимизовати. Такође, не постоји дељење параметара где се знање о једном делу секвенце не преноси на исти део секвенце у другом контексту. На пример, када је за предвиђање наредне речи у секвенци потребна информација која се налази много раније у секвенци или када је у контексту измењена само једна реч. Из тог разлога се користе специјлаизоване класе вештачких неуорнских мрежа као што су конволуционе и рекурентне вештачке неуорнске мреже.

2.5.3.2.3.1 Конволуционе вештачке неуронске мреже

Конволуционе вештачке неуронске мреже су првобитно развијене за решавање проблема из домена компијутерске визије, док су се тек недавно показале ефикасним за *NLP* задатке попут семантичког парсирања, моделовања реченица, класификације реченица (Kim, 2014). Стога ће архитектура *CNN* модела прво бити објашњена на примеру из домена компијутерске визије, након тога и на примеру из *NLP* домена.

Назив конволуционе вештачке неуронске мреже указује на то да мрежа користи математичку операцију која се назива конволуција. Модели овог типа мреже представљају регуларизовану верзију модела вишеслојног перцептрона (Секција 2.5.3.2.1) чији неурони у слојевима нису сви међусобно повезани. Скривени слојеви *CNN* модела укључују слој конволуције (енгл. *convolution layers*), након чега уобичајено следују слојеви удруживања (енгл. *pooling layers*), слојеви нормализације (енгл. *normalization layers*), слојеви потпуно повезане вештачке неуронске мреже (енгл. *fully*

connected layers). На Слика 39 је приказан пример архитектуре *CNN* модела за задатак класификације слика на десет класа. Тачније, у датом примеру се врши препознавање једноцифрених бројева на сликама и врши додела класе вредности препознатих бројева ($\{0,1,\dots,9\}$).



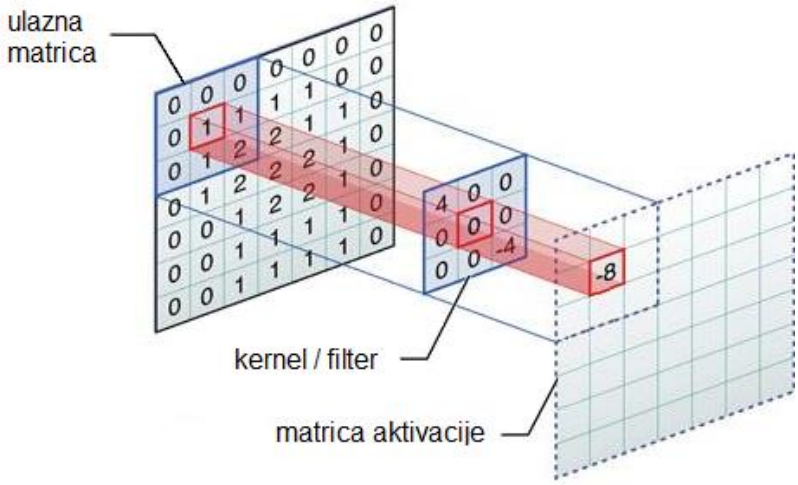
Слика 39 *CNN модел у домену компијутерске визије*[15]

Улаз *CNN* модела је матрица чије вредности представљају пикселе са слике. Слика може бити представљена као једнодимензионална матрица сиве нијансе (енгл. *grayscale*) и вишедимензионална матрица попут *RGB* модела (енгл. *Red, Green, Blue - RGB*). Први скривени слој мреже је слој конволуције који врши основну операцију ове мреже – конволуцију. Операција конволуције подразумева скаларни производ дела улазне матрице и филтера (кернела) одређене димензије. Циљ је извршити операцију конволуције над сваким делом улазне матрице.

Резултат једног корака конволуције је једна вредност на коју се додаје бијас и пропушта кроз активациону функцију (Секција 2.5.3.2.2.1). Начин рачунања једног корака конволуције је исти начину функционисања једног неурона (Секција 2.5.3.2.1). Стога се слој конволуције може посматрати као слој неурона који су повезани само са једним делом улазне матрице. Вредности филтера конволуције представљају тежине неурона а излазне вредности формирају матрицу активације. Међутим, сви неурони користе исте тежине, односно вредности филтера, те је стога обучавање знатно ефикасније у поређењу са вишеслојним перцептроном са истим бројем неурона.

Пример првог корака конволуције са филтером димензије 3x3 је дат на Слика 40.

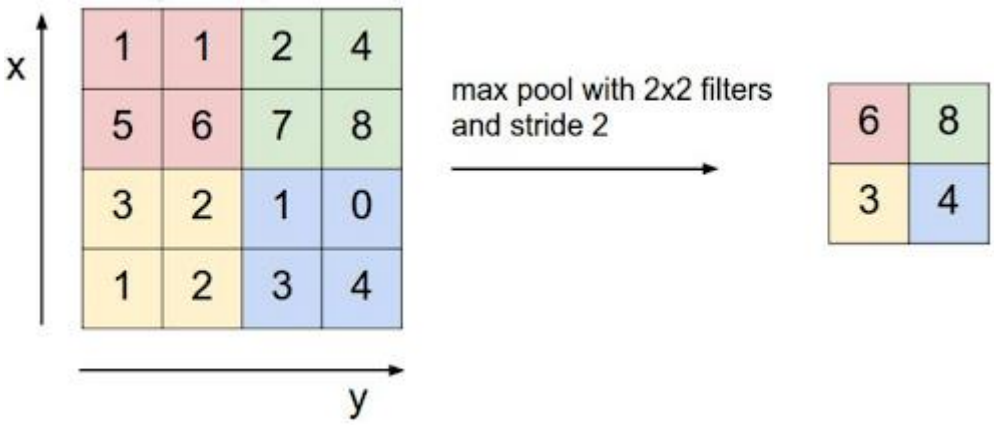
[15] Извор: https://github.com/alfo1995



Слика 40 *Конволуција филтером 3x3*

У наредним корацима се филтер помера за једну колону улазне матрице, рачунају активације и наставља даље до краја матрице. Након тога се филтер помера за један ред на ниже, враћа на почетак матрице и понавља се процес до краја доњег десног угла улазне матрице. Резултат слоја конволуције је матрица активација која садржи издвојене карактеристике слике. Величина матрице активације зависи од димензије филтера и величине корака. На пример, ако је димензија улазне матрице 32x32 а димензија филтера 5x5, са кораком један (примењеног на примеру изнад) се добије матрице активације димензије 28x28. Ако би корак био повећан на три, онда се добије матрице активације димензије 10x10. Док корак од два не би био примењив за дату димезију улазне матрице и филтера.

Применом више слојева конволуције се издвајају детаљније карактеристике улазне матрице. Међутим, применом више слојева конволуције са већим филтерима се димензије матрица активација смањују брзо што може довести до губитка информација. Из тог разлога се упоредо са кораком померања филтера конволуције примењује и метода испуњавања оквира улазне матрице нулама (енгл. *padding*). Циљ ове методе је да спречи или ублажи смањење димензија матрица активација и тиме сачува што више информација. Међутим, на крају *CNN* модела се углавном налазе слојеви потпуно повезане вештачке неуронске мреже те је стога циљ смањити број улазних података а при томе сачувати што више информација. У овом случају се након слоја конволуције на матрици активације примењује слој удруживања који ради на сличном принципу као слој конволуције. Користи се филтер одређене димензије који помера по матрици активације са одређеним кораком. При томе се уместо конволуције примењује функција за одређивање максимума, суме или просека. Пример слоја удруживања са филтером 2x2, кораком два и функцијом максимума је дат на Слика 41.

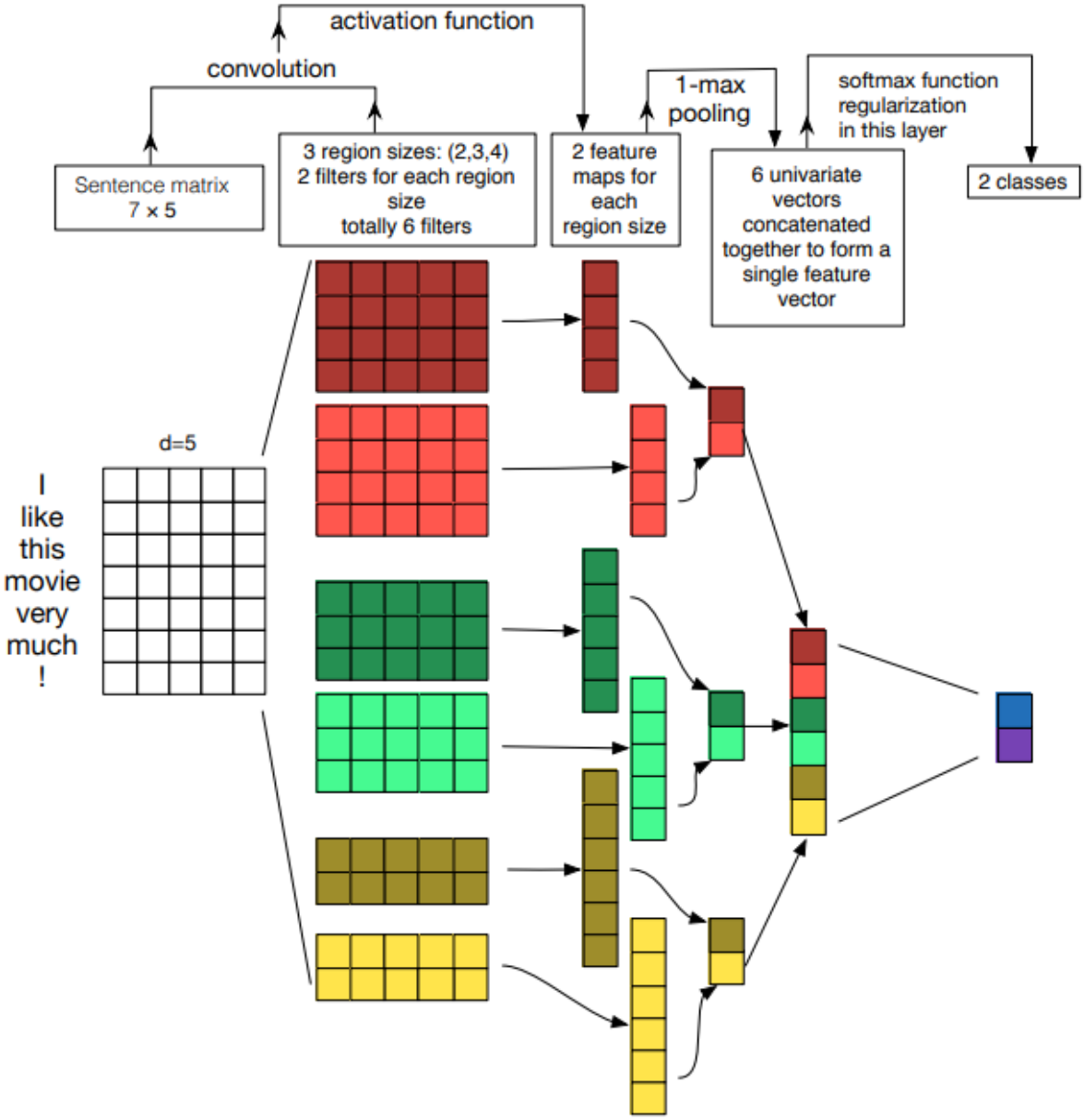


Слика 41 *Пример слоја удруживања са функцијом максимума*

Слојеви конволуције и удруживања се често понављају више пута у *CNN* моделу како би се издвојили што детаљније карактеристике (Слика 39). Добијене карактеристике се углавном прослеђују слојевима потпуно повезане вештачке неуронске мреже на основу које се може вршити класификација или детекција објеката на слици.

Конволуционе вештачке неуронске мреже у *NLP* области се одлично примењују за проналажење локалних језичких шаблона. Улази мреже могу бити произвољне величине на основу којих се могу издвојити смислени локални језички шаблони који су осетљиви на редослед речи. Ове мреже врло добро раде на проналажењу фраза или идиома до одређене дужине у дугим реченицама или документима. Уместо пиксела слике, улаз *CNN* модела за *NLP* задатке су реченице или документи који су репрезентовани матрицом. Сваки ред матрице најчешће одговара једном вектору речи добијених применом токенизатора претходно обучених језичких модела попут *Word2Vec*, *GloVe*, *ULMFiT*, *ELMo*, *BERT*, и тако даље. То би значило да ако улазна секвенца има седам речи које су репрезентоване вектором димензије пет, да је улазна матрица димензије 7x5.

У компијутерској визији се филтери превлаче преко локалних делова слике, док се у случају *NLP* задатака филтери уобичајено превлаче преко редова улазне матрице (речи). Стога је ширина филтера обично исте ширине као и улазна матрица. Док висина филтера варира и уобичајено обухвата две до пет речи у једном превлачењу (Zhang and Wallace, 2015). Пример *CNN* модела за *NLP* задатак класификације реченица аутора (Zhang and Wallace, 2015) је дат на Слика 42.



Слика 42 *Пример CNN модела за NLP задатак класификације реченица*

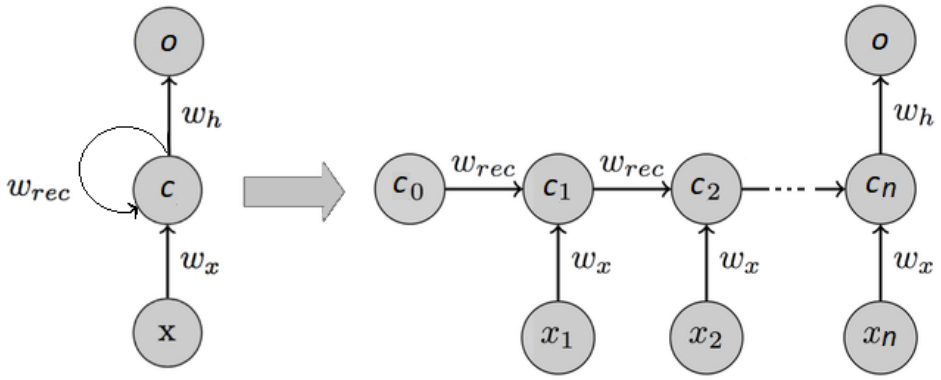
Применом филтера на улазну матрицу се врши конволуција и пропуштањем добијених вредности кроз активациону функцију се добијају вектори активација, односно карактеристика. Затим се на сваким вектором карактеристика примењује слој удруживања са функцијом максимума и формира један вектор карактеристика димензије броја примењених филтера. У крајњем слоју *CNN* модела се уобичајено за задатак класификације додаје потпуно повезана вештачка неуронска мрежа са софтмакс активационом функцијом (Секција 2.5.3.2.2.1.4) која одређује класе.

2.5.3.2.3.2 Рекурентне вештачке неуронске мреже

Нешто специјализованији модели за секвенцијалне податке попут текста су рекурентне вештачке неуронске мреже. Ове мреже узимају секвенцу података произвољне величине и производе вектор фиксне дужине тако да сумира улазну секвенцу. Међутим, рекурентне мреже се ретко користе као самосталне компоненте, него се чешће користе као компонента за обучавање која се надовезује на другу мрежну компоненту, формирајући један хибридни модел. Чест случај је да се на излаз рекурентне мреже надовежу улази мреже архитектуре потпуно повезаног вишеслојног перцептрона којом се покушава предвидети одређена вредност. У овом случају се рекурентна мрежа користи као модел за моделовање језика који трансформише улазне податаке у корисне векторске репрезентације за мрежу која је надовезана у наставку. Највећа предност рекурентних вештачких неуронских мрежа је могућност узимања у обзир дужих редоследа речи (секвенце) што доводи до импресивних предности у моделовању језика и предвиђању вероватноће наредне речи у секвенци. Из тог разлога су већина новијих истраживања управо усмерена ка рекурентним вештачким неронским мрежама.

Под појмом рекурентних вештачких неуронских мрежа се подразумевају све вештачке неуронске мреже које садрже цикличне везе између вештачких нурона. То значи да су одређени вештачки неурони или слојеви мреже, директно или индиректно, зависни од неког претходног излаза тј. стања. У зависности од *NLP* задатка, дужина улаза и излаза варира и стога постоји више типова рекурентних вештачких неуронских мрежа. На пример, у случају класификације се може применити архитектура са више улаза и једним излазом (енгл. *many-to-one*). Улаз у овом типу архитектуре модела најчешће представљају речи из једног сегмента текста (фразе, реченице, пасуса) а излаз вредност предвиђене класе.

За сваки вештачки неурон у рекурентној мрежи постоје два извора информација. Један је тренутни улаз а други је рекурентни улаз (повратни) који се може сматрати као један корак уназад кроз време. Са обзиром на то да рекурентна вештачка неуронска мрежа зависи од тренутног и претходних улаза онда се често каже да она садржи меморију. Захваљујући повратним спрегама кроз време је омогућено дељење параметара између вештачких неурона тј. повезивање знања у дугачким секвенцама. Пример овог типа архитектуре рекурентне вештачке неуронске мреже са једним слојем је дат на Слика 43. Структура рекурентне вештачке неуронске мреже (Слика 43, лево) се може представити у „развијеној“ форми дубоке вештачке неуронске мреже (Слика 43, десно) где рекурентна веза повезује претходни и тренутни вештачки неурон у времену.



Слика 43 *Пример структуре рекурентнтне вештачке неуронске мреже*

Стање вештачког неурона c_t у одређеном времену t зависи од улаза x_t и стања претходног вештачког неурона c_{t-1}, помножених одговарајућим тежинама. Формула рачунања стања вештачког неурона c_t је дата у наставку:

$$c_t = f(x_t \cdot w_x + c_{t-1} \cdot w_{rec} + b)$$

Где f представља функцију активације (најчешће функција хиперболичне тангенте), тежински параметар w_x тренутног улаза x_t, тежински параметар w_{rec} рекурентне везе стања претходног вештачког неурона c_{t-1} и b бијас параметар. Током обучавања модела рекурентне вештачке неуронске мреже стање c_t евалуира на основу тренутног улаза x_t и стања претходног вештачког неурона c_{t-1}, што даље утиче на евалуацију стања наредног вештачког неурона c_{t+1}. На овај начин се вештачки неурони уланчавају (Слика 43, десно) и тиме формирају меморију модела кроз време. Излаз класификационог модела o се добија множењем последњег стања c_n и тежинског параметра w_h и пропуштањем кроз функцију активације (најчешће софтмакс). У случају архитектуре са више улаза и излаза (енгл. *many-to-many*) се излаз модела o може добити након сваког корака тј. стања c_t.

Међутим, овакв структура вештачке неуронске мреже не решава проблем дужег контекста у потпуности. Услед веома дугачких улазних секвенци (контекста) настаје проблем нестајућег градијента. Последица овог проблема је губитак информација о ранијим стањима (кроз време) тј. немогућност чувања информација веома дугачких секвенци. Решење овог проблема су специфичне архитектуре рекурентних мрежа попут модела дуготрајне-краткотрајне меморије (енгл. *Long Short-Term Memory – LSTM*) и рекурентне јединице са механизмом затварања (енгл. *Gated Recurrent Units – GRU*). У оба случаја се вештачки неурон рекурентне мреже замењује комплекснијом јединицом која садржи више операција над подацима. Такође, уводе концепт капија (енгл. *gates*) којим се контролише измена стања скривеног слоја и тиме спречава проблем нестајућег градијента.

Јединица *LSTM* модела (Hochreiter and Schmidhuber, 1997) проблем дугачког контекста решава из два корака. У првом кораку врши уклањање информација које нису више потребне у датом контексту а у другом кораку додаје информације које ће са већом вероватноћом бити потребне за успешније моделовање контекста. Дакле, пре измене стања *LSTM* јединице се врши провера релевантности информације за цео контекст. Концепт капије у оквиру јединице је имплементиран вештачким неуронима са сигмоидалном функцијом активације. *LSTM* модел поред стања јединице уводи и скривено стање јединице које се узима у обзир при измени тренутног стања јединице. Процес измене стања *LSTM* јединице се може дефинисати следом следећих једначина:

$$\hat{c}_t = \tanh(x_t \cdot w_x + h_{t-1} \cdot w_{rec} + b_c)$$

$$g_f = \sigma\big(x_t \cdot w_x + h_{t-1} \cdot w_f + b_f\big)$$

$$g_u = \sigma(x_t \cdot w_x + h_{t-1} \cdot w_u + b_u)$$

$$g_o = \sigma(x_t \cdot w_x + h_{t-1} \cdot w_o + b_o)$$

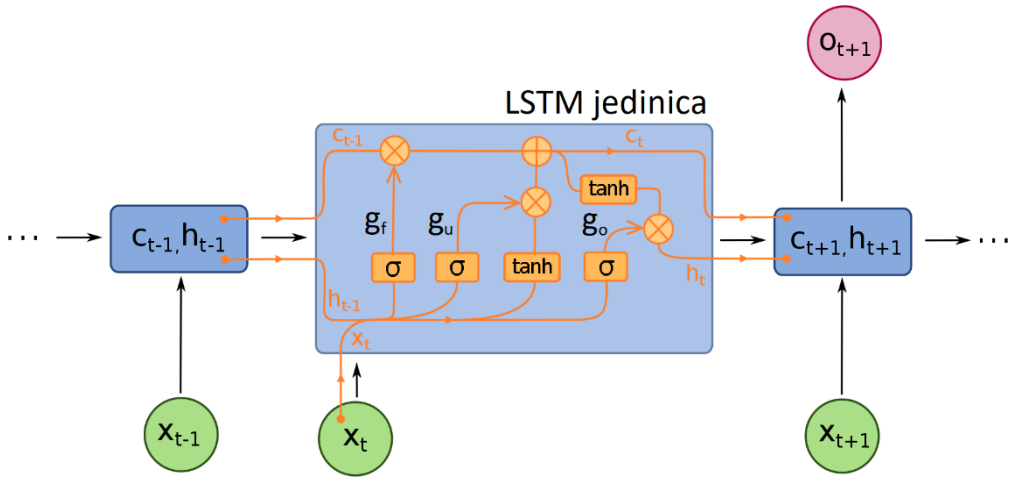
$$c_t = g_u \cdot \hat{c}_t + g_f \cdot c_{t-1}$$

$$h_t = g_o \cdot \tanh(c_t)$$

LSTM јединица на улаз добија стање претходне јединице c_{t-1} и скривено стање претходне јединице h_{t-1}. Рачунање тренутног стања *LSTM* јединице $\hat{c}_t$ се своди на исти принцип као и код стандардне рекурентне вештачке неуронске мреже. Тренутно стање $\hat{c}_t$ представља кандидата за измену крајњег стања c_t у зависности од вредности капија.

Прво се рачуна вредност капије g_f за брисање информација из контекста c_{t-1} које више нису потребне. Вештачки неурон капије g_f рачуна тежинску суму претходно скривеног стања и тренутног улаза и пропушта кроз сигмоидалну функцију активације. Након тога се на сличан начин рачуна вредност капије g_u за додавање информација из тренутног контекста $\hat{c}_t$. Вредност крајње капије g_o служи за доношење одлуке које информације су потребне, односно корисне, за тренутно скривено стање h_t. Вредности свих капија су у опсегу $(0, 1)$ што утиче на множиоце тако што пропушта вредности стања у већој или мањој мери. Тачније, имају ефекат маскирања контекста тако што се одређене информације задржавају или одбацују.

Коначно тренутно стање јединице c_t се добија маскирањем стања претходне јединице c_{t-1} и кандидата стања тренутне јединице $\hat{c}_t$ одговарајућим вредностима капија g_f и g_u. Вредношћу крајње капије g_o се маскира добијено коначно тренутно стање јединице c_t, пропуштене кроз тенгентну функцију активације, чиме се формира скривено стање тренугне јединице h_t. Графички приказ описаног поступка рада *LSTM* јединице је дат на Слика 44.



Слика 44 *LSTM јединица*

Хоризонталним уланчавањем више *LSTM* јединица се формира модел који је врло ефикасан у меморисању веома дугачких секвенци (контекста). Међутим, његова комплексност захтева више података и дуже обучавање што у одређеним случајевима није примењиво. Из тог разлога су настале *GRU* јединице (Cho *et al.*, 2014) које такође раде на концепту капија али су мање комплексне, омогућавају брже обучавање и веома су ефикасне у одређеним случајевима. У поређењу са *LSTM* јединицом, *GRU* јединица не садржи излазну капију и тренутно стање јединице. Тачније, *GRU* јединица не врши измену скривеног стања одвојено од тренутног стања јединице, него се скривено стање једино размењује међу *GRU* јединицама. Ради конвенције са уводним делом рекурентних вештачких неуронских мрежа и *LSTM* јединицом, скривено стање *GRU* јединице се означава са c_t. Процес измене стања *GRU* јединице је поједностављен и може се дефинисати следом следећих једначина:

$$\hat{c}_t = \tanh(x_t \cdot w_x + g_r \cdot c_{t-1} \cdot w_{rec} + b_c)$$

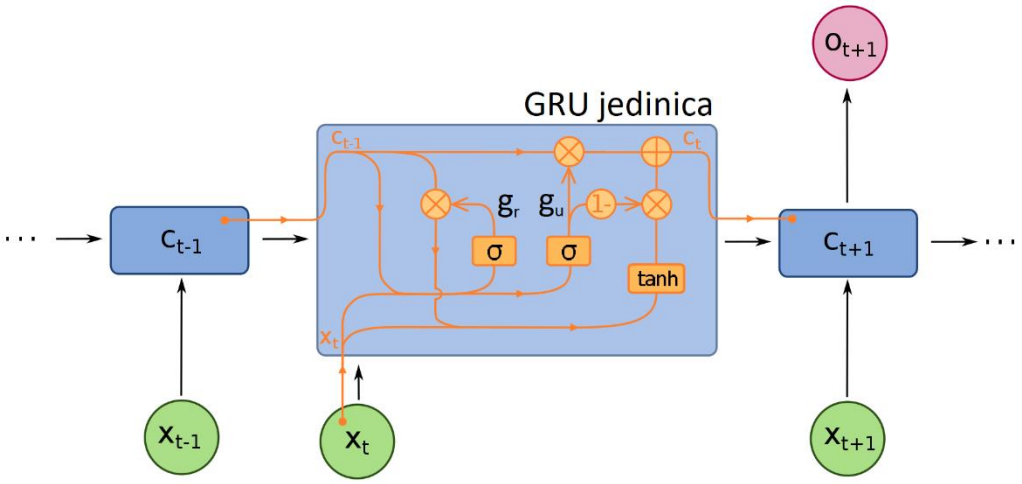
$$g_r = \sigma(x_t \cdot w_x + c_{t-1} \cdot w_r + b_r)$$

$$g_u = \sigma(x_t \cdot w_x + c_{t-1} \cdot w_u + b_a)$$

$$c_t = g_u \cdot \hat{c}_t + (1 - g_u) \cdot c_{t-1}$$

Рачунање тренутног стања *GRU* јединице $\hat{c}_t$ се своди на исти принцип као и код *LSTM* јединице са једном изменом. Стање претходне *GRU* јединице c_{t-1} се маскира вредношћу капије релевантних информација g_r. Вредност g_r капије одређује колико је стање претходне *GRU* јединице c_{t-1} релевантно за рачунање тренутног стања $\hat{c}_t$. Тренутно стање $\hat{c}_t$ представља кандидата за измену крајњег стања c_t у зависности од вредности g_u капије. Капија g_u служи за додавање информација из тренутног контекста $\hat{c}_t$ а уједно и уклањање информација из стања претходне *GRU* јединице c_{t-1}. На основу

маске капије g_u се формира коначно стање *GRU* јединице c_t. Графички приказ описаног поступка рада *GRU* јединице је дат на Слика 45.



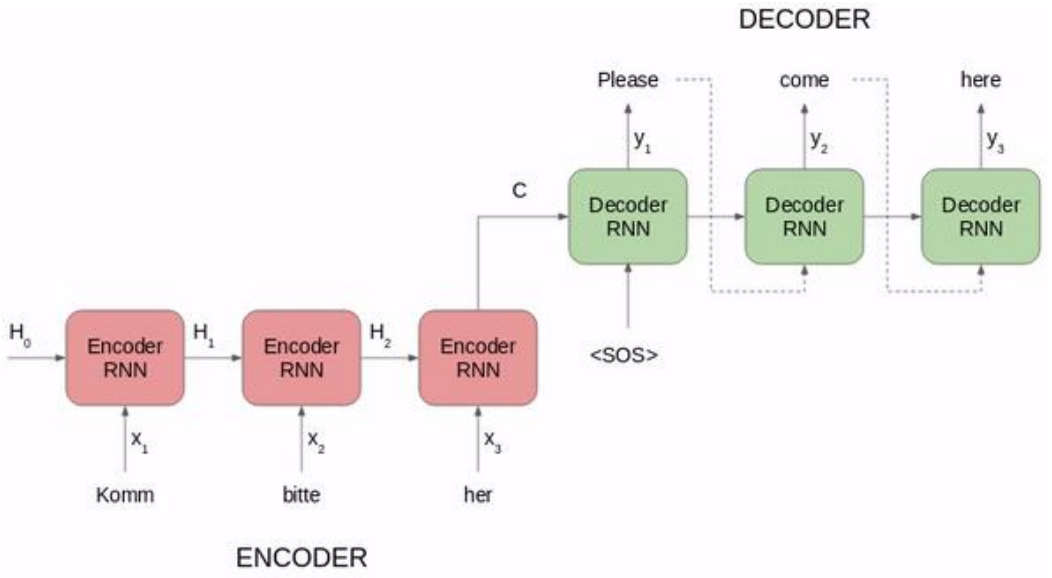
Слика 45 *GRU јединица*

Вештачке јединице *GRU* и *LSTM* модела се могу повезивати у ланац тј. слојеве, чиме се повећава способност мреже да научи комплексније репрезентације података на различитим нивоима апстракције кроз слојеве мреже. Примери популарних архитектура *RNN* мреже су дати у наставку.

2.5.3.2.3.3 Архитектура енкодер-декодер

У развоју архитектура вештачких неуронских мрежа је веома битна архитекрура са више улаза и излаза (енгл. *many-to-many*). Улаз у овом типу архитектуре модела најчешће представљају секвенцу речи из једног сегмента текста (фразе, реченице, пасуса) на једном језику а излаз секвенца речи на другом језику. Овај тип архитектуре се најчешће користи за *NLP* задатке попут машинског превођења, сумирања текста, аутоматско одговарање на питања, и тако даље. Овај тип архитектуре се често назива и „секвенца-у-секвенцу“ модел (енгл. *sequence-to-sequence - seq2seq*).

seq2seq архитектура са *RNN* јединицама (*GRU*, *LSTM*) је најпре коришћена за сврху машинског превођења. Ова архитекрура се састоји из два дела, две *RNN* мреже, енкодера и декодера. На основу улазне секвенце енкодера се формира вектор контекста који се прослеђује декодеру на основу које се формира излаз. Улазну секвенцу представљају речи на једном језику, док излазна секвенца представљају речи на другом језику. Пример *seq2seq* архитектуре са *RNN* јединицама за задатак превођења секвенци са немачког на енглески језик је дат на Слика 46.



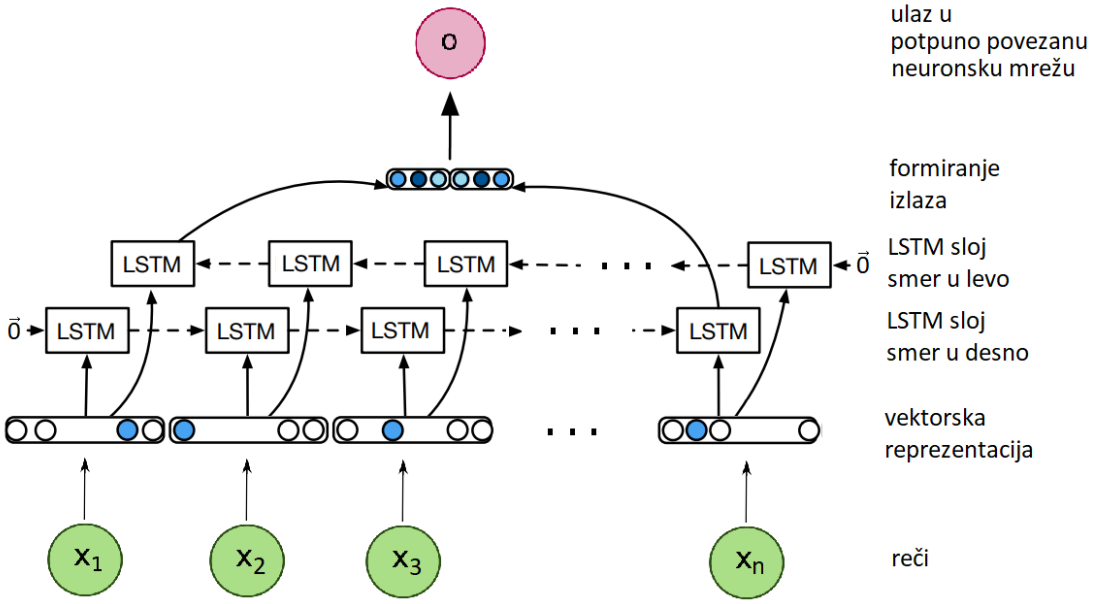
Слика 46 *Архитектура енкодер-декодер*

У сваком временском тренутку t енкодер узима векторе речи x_t из улазне секвенце и скривено стање H_t из претходног корака. Скривено стање H_t се допуни у сваком тренутку t. Скривено стање из последње *RNN* јединице енкодера садржи информације о улазној секвенци и назива се вектор контекста. Затим се вектор контекста прослеђује декодеру на основу које се генерише излазна секвенца y_t.

Архитектура енкодер-декодер се у пракси показала као велики помак на задатку машинског превођења. Међутим, ограничење ове архитектуре је дужина секвенци. У веома дугачким секвенцама долази до губитка контекста јер *RNN* мрежа не успева да задржи све потребне информације. Такође, са практичне стране, процес обучавања овакве архитектуре није могуће паралелизовати што утиче на време развоја модела на великим корпусима.

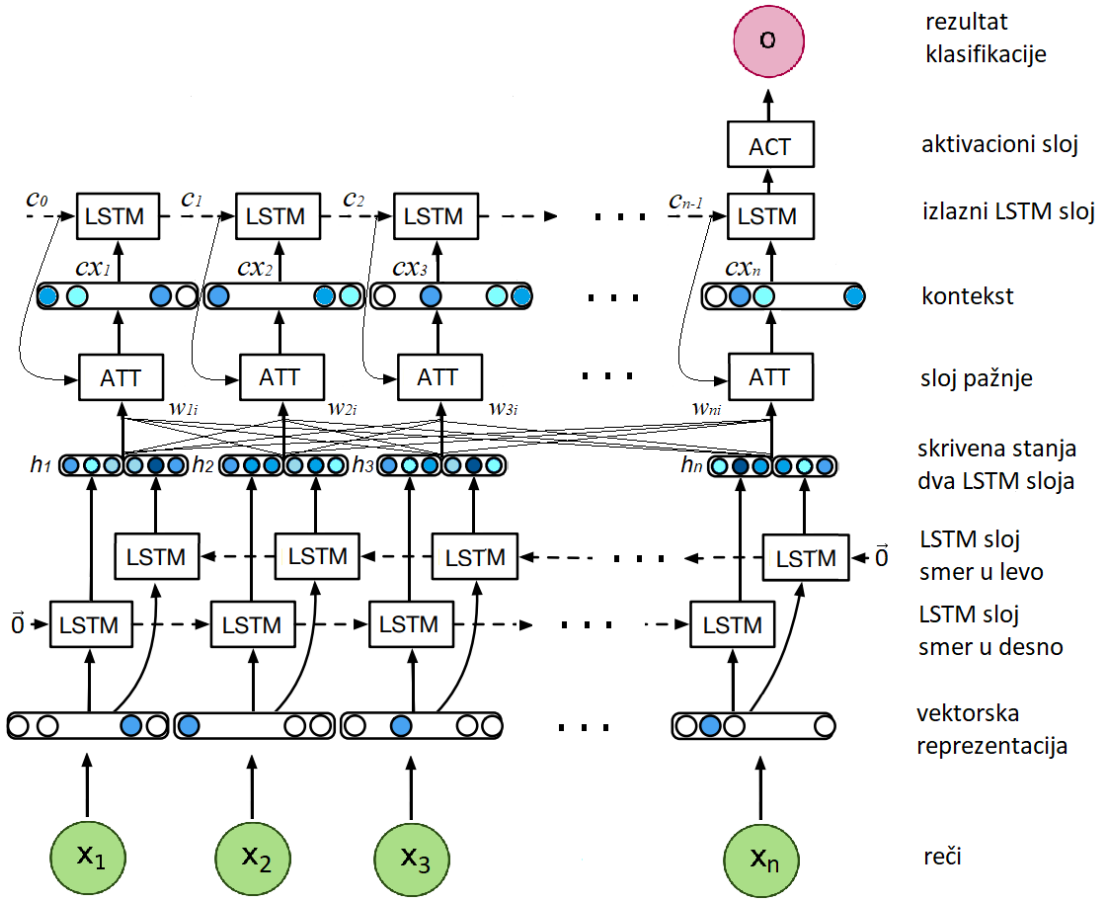
2.5.3.2.3.4 Ахитектура слоја пажње

Архитектура рекурентне вештачке неуронске мреже која садржи два смера се назива двосмерна *RNN* мрежа (енгл. *bidirectional RNN*). Овај тип архитектуре се састоји од једног слоја хоризонтално уланчаних комплексних *RNN* јединица (*GRU, LSTM*) повезаних у смеру од првог до последњег улазног податка и другог слоја хоризонтално уланчаних јединица повезаних у супротном смеру, од последњег до првог улазног податка секвенце. На овај начин се узимају у обзир информације из оба смера чиме се даје подједнак значај (утицај) сегментима текста са почетка и краја секвенце. Графички приказ архитектуре овог примера мреже је дат на Слика 47.



Слика 47 *Пример архитектуре двосмерног LSTM модела*

Проблем двосмерних *RNN* модела је губљење релевантних информација у дугачким секвенцама јер се иста пажња посвећује свакој речи у секвенци. Из тог разлога се на излаз рекурентне вештачке неуронске мреже додаје још један слој тежина којим се додељује већа пажња одређеним деловима секвенце (енгл. *attention mechanism*). Овим механизмом пажње се омогућава моделу да директно приступи стању *GRU* или *LSTM* јединице у било којој позицији секвенце и додели им одређену тежину важности за тренутни улаз. Излазни вектор слоја пажње се најчешће надовезује на још један слој рекурентне вештачке неуронске мреже која постиже знатно боље перформансе у *NLP* задацима у односу на моделе без слоја пажње. Пример архитектуре двосмерног *LSTM* модела са слојем пажње за задатак класификације је дат на Слика 48.



Слика 48 *Пример архитектуре двосмерног LSTM модела са слојем пажње*

У датом примеру се векторске репрезентације речи x_t, где $t \in (1, n)$, прослеђују одговарајућим *LSTM* јединицама у *LSTM* слоју у оба смера, чија скривена стања се групишу у један вектор h_t. Сваки вектор h_t садржи информације о целој улазној секвенци речи са пажњом на делове који окружују t-ту реч у улазној секвенци. У слоју пажње се за сваку реч t рачуна контекст cx_t као тежинска сума вектора скривених стања h_t и одговарајућих тежна пажње w_{ti} за позицију i, где $i \in (1, n)$, по следећој формули:

$$cx_t = \sum_{i=1}^{n} h_i w_{ti}$$

Тежна пажње w_{ti} се за сваки вектор скривених стања h_t рачуна применом софтмакс функције (Секција 2.5.3.2.2.1.4) на следећи начин:

$$w_{ti} = \frac{e^{p_{ti}}}{\sum_{j=1}^{n} e^{p_{tj}}}$$

Где је сума тежна пажње w_{ti} за реч t једнака јединици ($\sum_{i=1}^{n} w_{ti} = 1$). Где p_{ti} представља меру поравњања којом се оцењује колико добро се подударају улази око

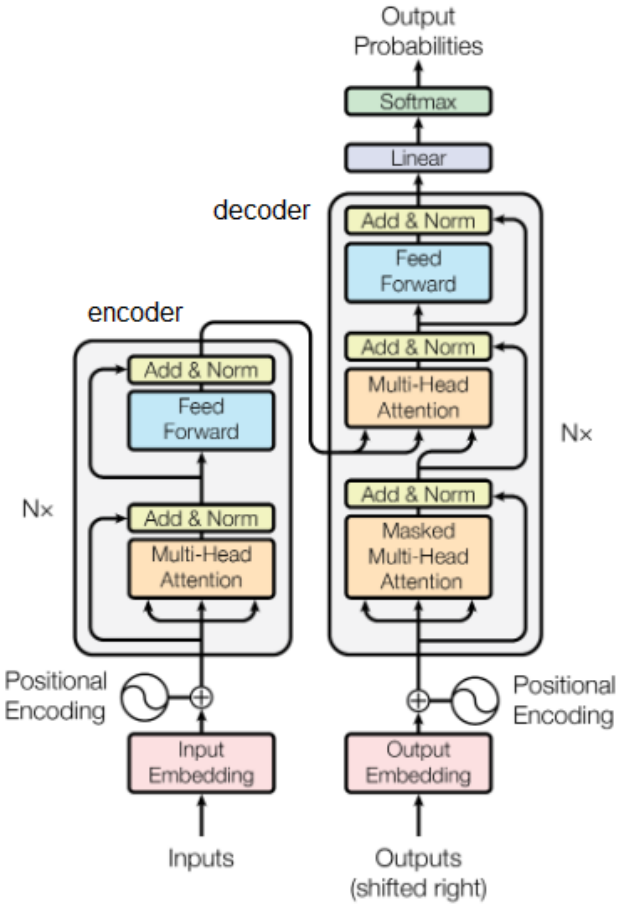
положаја i и излази на положају t. Ова мера се одређује обучавањем неуронске мреже са једним слојем nn чији су улази претходно скривено стање излазног *LSTM* слоја c_{t-1} и h_i вектор скривених стања i-те речи улазне секвенце на следећи начин:

$$p_{ti} = nn(c_{t-1}, h_i)$$

2.5.3.2.3.5 Архитектура трансформера

Компанија Гугл је 2017. предложила архитектуру трансформера (Vaswani *et al.*, 2017) како би решили проблеме секвенцијалних модела попут машинског превођења. До тада актуелни модели за превођење језика су углавном били засновани на *RNN* имеплементацијама енкодер-декодер архитектуре. У пракси се показало да је обучавање *RNN* модела споро и да је због њихове архитектуре онемогућена паралелизација обраде података. Стога је архитектура трансформера заснована на слоју пажње али изоставља *RNN* имплементацију.

Трансформер архитектура је базирана на енкодер-декодер структури са механизмима пажње без рекурентне секвенцијалне обраде података која је обавезна код рекурентних вештачких неуронских мрежа. Додавањем више слојева пажње се добија вишеслојни механизам пажње којим се омогућава посвећивање пажње на више контекста у једној истој секвенци текста. Погодност коју трансформер архитектура доноси је могућност паралелизације процеса обучавања без потребе обраде података по реду (што је случај са рекурентним моделом). Тиме је скраћено време обучавања и повећане могућности учења на већим скуповима података. Архитектура трансформера је приказана на Слика 49 (Vaswani *et al.*, 2017) и у наставку објашњена на задатку машинског превођења са српског на енглески језик.



Слика 49 *Архитектура трансформера*

Задатак машинског превођења се заснива на предвиђању излазне секвенце на једном језику (превода) на основу дате улазне секвенце на другом језику. Приликом обучавања модела се речи улазне и излазне секвенце трансформишу у векторске репрезентације где се додају информације о позицији речи у секвенци. С обзиром на то да овај модел не садржи рекурентну везу или конволуцију (попут *RNN* и *CNN*), позиционо енкодовање омогућава моделовање контекста. Сумарне векторске репрезентације речи, улазне секвенце и позиционог енкодовања се прослеђују у енкодер, а излазне секвенце и позиционог енкодовања у декодер део.

Енкодер део се састоји из две компоненте, компоненте само-пажње и компоненте потпуно повезане вештачке неуронске мреже са два слоја. У компоненти само-пажње се за сваку реч формира вектор пажње (контекста) који показује колико је дата реч релевантна за остале речи у секвенци. На пример, вектори пажње улазне секвенце „*Danas je divan dan*“ је дат на Слика 50.

Danas:	[Danas je divan dan]	-> [0,5 0,3 0,1 0,1]
je:	[Danas je divan dan]	-> [0,2 0,8 0,2 0,1]
divan:	[Danas je divan dan]	-> [0,1 0,1 0,5 0,3]
dan:	[Danas je divan dan]	-> [0,3 0,1 0,2 0,5]

Слика 50 *Пример вектора пажње речи улазне секвенце*

На примеру се може видети да је за сваку посматрану реч највећа пажња (вероватноћа) усмерена управо на исте речи у секвенци. Ове вредности не доносе информације о контексту него вероватноће осталих речи у секвенци. Из тог разлога су аутори предложили коришћење вишеструке пажње (енгл. *Multi-head Attention*) тј. механизам формирања више вектора пажње за сваку реч и преузимање по једног вектора на основу тежинског просека. У наставку се вектори пажње сваке речи пропуштају кроз компоненту потпуно повезане вештачке неуронске мреже и трансформишу у форму која одговара наредној компоненти.

Декодер део такође садржи поменуте две компоненте у енкодер делу са додатном компонентом маскиране само-пажње на почетку декодер дела. Компонента маскиране само-пажње за сваку реч излазне секвенце формира векторе пажње при чему се маскирају речи које се налазе после посматране речи у секвенци. На тај начин се омогућава бољи квалитет обучавања модела за дати задатак. На пример, вектори пажње речи излазне секвенце „*Today is a beautiful day*“ су дати на Слика 51.

Today:	[**Today** is a beautiful day]	-> [0,1 0,0 0,0 0,0 0,0]
is:	[Today is a beautiful day]	-> [0,2 0,8 0,0 0,0 0,0]
a:	[Today is a beautiful day]	-> [0,1 0,2 0,7 0,0 0,0]
beautiful:	[Today is a beautiful day]	-> [0,1 0,3 0,1 0,5 0,0]
day:	[Today is a beautiful day]	-> [0,1 0,2 0,1 0,1 0,6]

Слика 51 *Пример вектора пажње речи излазне секвенце*

На примеру се може видети да се пажња усмерава само на претходне речи од посматране речи док су вероватноће за остале речи нуле. Затим се вектори пажње речи излазне секвенце заједно са векторима пажње речи улазне секвенце прослеђују наредној компоненти декодера – компоненти само-пажње. У овој компоненти се врши формирање заједничких вектора пажње речи улазне и излазне секвенце који одређује у каквој су релацији речи обе секвенце. Тачније, у овом кораку се врши мапирање речи између језика где се усмерава пажња на оне речи које највише утичу на очекивани исход. Излаз ове компоненте су појединачни вектори пажње за сваку реч улазне и излазне секвенце. Ови вектори пажње се даље пропуштају кроз компоненту потпуно повезане вештачке неуронске мреже и трансформишу у форму која одговара наредној компоненти. На излаз декодера се додаје још једна потпуно повезана вештачка

неуронска мрежа која на основу појединачних вектора пажње формира излазну секвенцу вероватноћа која одговара димензији речника језика излазне секвенце (превода). Крајњи резултат су предвиђене наредне речи речи излазне секвенце.

Компонента вишеструке пажње је детаљније приказана на Слика 52 (Vaswani *et al.*, 2017) и описана у наставку. За сваку реч у секвенци се формирају три вектора - упит *Q* (енгл. *Query – Q*), кључ *K* (енгл. *Key – K*) и вредност *V* (енгл. *Value – V*), која издвајају три различите компоненте вектора речи. С обзиром на то да вектори *Q*, *K* и *V* потичу од истог вектора речи, овај механизам се и назива само-пажње (енгл. *Self-Attention*). Циљ је пронаћи значајне речи *K* из секвенце за сваку посматрану реч *Q*. Ово се постиже пропуштањем производа вектора *Q* и *K* кроз софтмакс функцију (Секција 2.5.3.2.2.1.4) и множењем са вектором вредности *V*. Производ вектора *Q* и *K* се дели скалирајућим фактором ради смањења великих вредности. Функција пажње (Слика 52, десно) која се користи за обучавање мапирање вектора упита *Q* и парова вектора кључа *K* и вредности *V* је описана формулом у наставку:

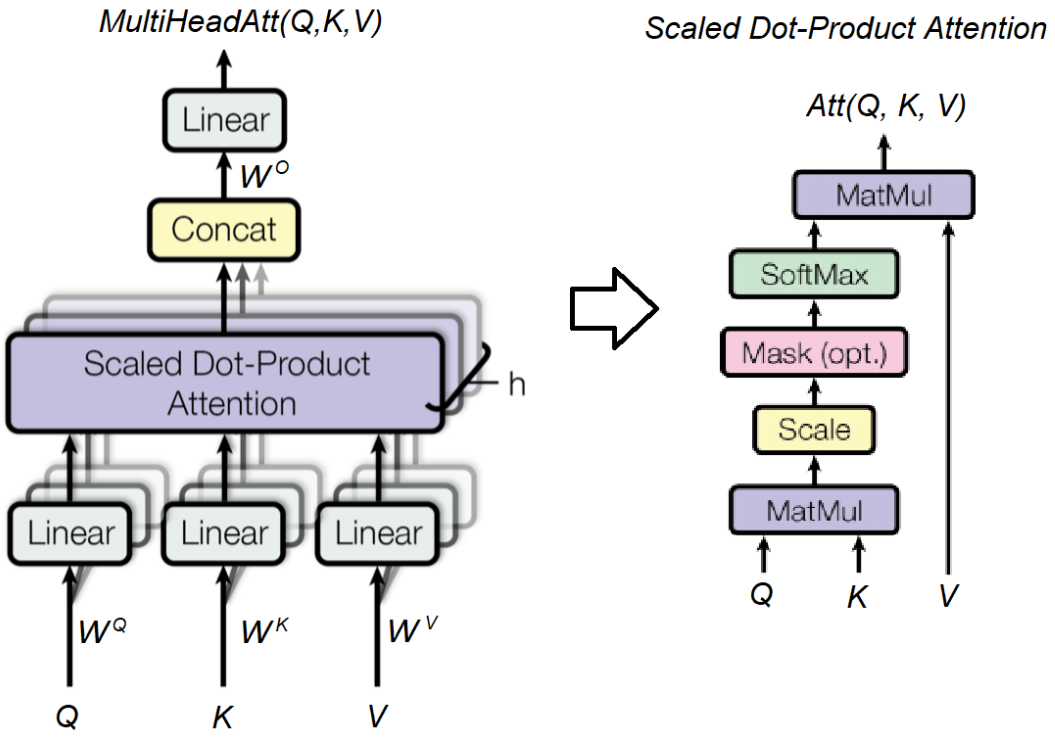
$$Att(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{\dim\,(K)}}\right)\cdot V$$

Умножавањем претходног поступка се формирају вишеструки вектори пажње за једну реч (Слика 52, лево). Крајњи вектор пажње речи се добија на основу тежинског просека свих вектора пажње на следећи начин:

$$MultiHeadAtt(Q,K,V) = Concat(h_1,h_2,\dots,h_n)\cdot W^O$$

$$h_i = Att_i\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

У оквиру сваке компоненте енкодера и декодера се налази слој нормализације који служи за стабилизацију обучавања, односно избегавање проблема нестајућег градијента (Секција 2.5.3.2.2.1.1). Наиме, све компоненте вектора пажње у обучавајућем скупу (или подскупу) се нормализују независно тако да средња вредност компоненти вектора буде близу нуле а стандардна девијација близу јединице.



Слика 52 *Компонента вишеструке пажње трансформера*

3 Преглед актуелног стања у области

Почетна истраживања у области сентимент анализе су била усмерена на анализи текста на нивоу документа тј. целог коментара (Dave *et al.*, 2003; Pang *et al.*, 2002; Pang and Lee, 2005; Tang, 2015; Tang and Liu, 2015; Turney, 2002; Xu *et al.*, 2016). Међутим, сентимент анализа на нивоу документа је мање ефективна ако је потребна детаљнија анализа (Thet *et al.,* 2010). На пример, рецензија из домена филмова може да се односи на глумце са позитивним сентимент поларитетом док помињање радње филма може бити у негативном контексту. Такође је могуће да једна реченица садржи више различитих сентимент поларитета. На пример, „Главни глумац филма је одличан али радња филма делује неповезано.“. Из потребе да се разумеју специфични сентименти за различите аспекте одређеног ентитета настала је област аспектно базиране сентимент анализе (Секција 2.3), као под област сентимент анализе (Секција 2.2).

У наставку су описана *ABSA* истраживања из општег домена и домена високог образовања. Прво су описани *ABSA* приступи засновани на правилима и речницима, након чега су описани *ABSA* приступи засновани на алгоритмима машинског учења. На крају овог поглавља су описана истраживања из области сентимент анализе за српски језик.

3.1 ABSA заснована на речницима и правилима

Већина почетних истраживања из области аспектно базиране сентимент анализе су се ослањала на постојеће алате и ресурсе за обраду природног језика попут ручно израђених правила и лексикона (Hu and Liu, 2004;Popescu and Etzioni, 2007; Thet *et al.,* 2010; Yi *et al.,* 2003). Приступи засновани на сентимент речницима се ослањају на примену унапред дефинисаних сентимент речника и агрегацију сентимент резултата у опште позитиван или негативан сентимент. Према речима аутора (Liu, 2015), позитивне речи и фразе исказују се жељена стања, док се негативним исказују нежељена. Скуп ових речи и фраза чине речник мишљења или сентимента. Аутор истиче да приступи

засновани на речницима поред речника сентимент речи и фраза такође укључују правила обраде језичких конструкција (нпр. „мењачи сентимента“, „али“ клаузуле) и типова реченица. Носиоци сентимент поларитета су објекти тј. аспекти који су овом кораку познати, дати или претходно идентификовани.

Најједноставнији алгоритам примене сентимент речника, који по аутору (Liu, 2015) остварују прилично добре резултате у пракси, се састоји из следећа четири корака, ондосно правила. Уколико се аспектна анализа врши на нивоу реченица, први корак је да се означе сентимент изрази (речи, фразе) у случају да постоји један или више аспеката (укључујући и ентитете). Сваком позитивном изразу се додељује резултат један (+1), док се негативном изразу додељује резултат минус један (-1). На пример, ако се посматра реченица „Предмет је веома занимљив али недостаје уџбеник на нашем језику.“, први сентимент израз („Предмет је веома занимљив“) има резултат један (+1) а други сентимент израз („недостаје уџбеник“) минус један (-1). У другом кораку се анализирају мењачи сентимента, односно речи негације који мењају сентимент поларитет. Речи попут не, ни, никада, ништа, нико и нигде су речи које мењају сентимент израза који се налази после њих. На пример, израз „ја могу то“ је позитивне оријентације због глагола „моћи“, док уз присуство речи негације „ја не могу то“ мења сентимент оријентацију на негативан поларитет. Трећи корак алгоритма примене сентимент речника се односи на случај анализе мишљења на нивоу реченице и појаву везника „али“. Везник „али“ има често супротну сентимент конотацију и означава да се сентимент дела реченице који следује разликује од сентимента дела реченице који претходи. Овде се могу анализирати и други везници типа „а“, него“ и „већ“. Последњи, четврти, корак подразумева функцију агрегације сентимент резултата добијеног у претходним корацима. Агрегација сентимента зависи и од нивоа анализе сентимента, те се за ниво реченице мора прво одредити позиције аспеката у реченици (ако их је више) и онда извршити агрегација сентимент резултата у односу на синтаксичке релације сентимент поларитета и аспекта. Синтаксичке релације се добијају синтаксичким парсирањем, где се у односу на врсте речи може одредити припадност сентимент поларитета. У случају нижег нивоа анализе попут сегмента реченице (клауза и фраза) се агрегирана сентимент оријентација односи на цео сегмент реченице.

Према начину формирања сентимент речника, аутор (Liu, 2015) је груписао приступе у следеће три групе. Прву групу чине ручни приступи који су напорни и временски захтевни. Ови приступи су првобитно коришћени за анализу сентимента док се касније углавном користе за проверу речника формираног аутоматским путем. Друга два приступа су заснована на речницима и корпусу. Приступ заснован на речницима користи постојеће речнике синонима и антонима и технику генерисања речи на основу речника синонима и антонима. На основу прикупљеног скупа позитивних и негативних речи се претражују постојећи језички речници за њихове синониме и антониме. Нове

пронађене речи синонима и антонима се додају у првобитне скупове позитивних и негативних речи и процес се понавља докле год не буде ни један синоним и антоним пронађен. На крају овог процеса је потребно ручно извршити проверу добијених речи и уклонити погрешне. Приступ заснован на корпусу је мање ефикасан него приступ заснован на речницима. Ови приступи се корсите за идентификацију доменско специфичних сентимент речи и адаптацију сентимент речника опште намене на одређени домен, користећи доменски корпус. Овај процес је компликован јер доменски специфични речник није довољан, где једна реч у одређеном контексту може имати позитивну оријентацију а у другом контексту негативну оријентацију.

Један од првих приступа заснованих на речницима и правилима (Yi *et al.* 2003) се базира на издвајању позитивног и негативног сентимента за одређене аспекте из рецензија о дигиталним камерама и музичким производима. На пример, за домен дигиталних камера издвајани су аспекти делова дигиталних камера попут објектива, батерије или меморијске картице, затим њихових атрибута попут цене, величине или трајања батерије. Док би за домен музике аспекти били песма, албум, текст, бенд итд. Аутори су користили ручно направљена правила базираним на синтаксичком парсирању како би у рецензијама идентификовали реченице које садрже одређени аспект. У оквиру реченице су прво издвајали именичке фразе које су након тога филтриране на основу инверзне фреквенције појављивања у целом корпусу. Издвојени аспекти из реченица су ручно прегледани и евалуирани од стране аутора и пријавили су просечну прецизност од 0,97 за аспект дигиталне камере и 1,00 за аспект музике. Сентимент поларитет за сваку реченицу је одређиван применом сентимент лексикона сачињеног од 2.500 придева и 500 именица. Како би одредили ком аспекту припада одређен сентимент, аутори су дефинисали правила (шаблоне) за препознавање фразе предиката реченице. Правила имају форму регуларних израза (Секција 2.5.2) где почетак шаблона чини глагол, крај аспект а између њих сентимент изрази које углавном чине придеви. На основу сентимент поларитета придева аутори су вршили доделу сентимента објекту реченице тј. аспекту. За задатак издвајања позитивног и негативног сентимента аутори су саопштили просечну прецизност од 0,87 и просечан одзив од 0,56.

У другом истраживању аутори (Hu and Liu, 2004) су се ослањали на фреквентне именичке фразе добијене *POS* таговањем. Аутори су анализирали сентимент у рецензијама о електронским производима са *Amazon*[16] и *Cnet*[17] сајта. Задатак је био да се идентификују аспекти електронских производа (камера, телефон, ДВД и мп3) и да се додели сентимент поларитет. Прво су *POS* таговањем у свакој реченици корпуса

[16] Сајт: www.amazon.com
[17] Сајт: www.cnet.com

означили именице и именичке фразе, које су затим филтриране применом асоцијативних правила. Употребом алгоритма за проналажење фреквентних подскупова из области анализе асоцијативних правила аутори су издвојили најфреквентније фразе које се појављују у један посто реченица корпуса а одстрањене оне које се не појављују у више од три реченице или су део већ неке друге фразе. Евалуацију задатка издвајања аспеката електронских производа аутори су извршили на основу ручно анотираног корпуса и постигли просечан одзив од 0,67 и прецизност од 0,79. Након тога, аутори су издвојили придеве из реченица које су користили за сентимент анализу. За одређивање сентимент поларитета су коришћени ручно формирани сентимент лексикони. Лексикони су сачињени од позитивних и негативних придева који су проширени синонимима и сличним придевима добијених на основу језичког ресурса мреже речи (*WordNet*[6]). Аутори су постигли просечну тачност од 0,64 и просечан одзив од 0,69, узимајући у обзир само позитивне и негативне сентименте.

Popescu and Etzioni (2007) су побољшали приступ аутора (Hu and Liu, 2004) тако што су увели софистицираније начин филтрирања именичких фраза везаних за аспекте електронских производа. Метода филтрирања се ослања на *PMI* информације између два термина (енгл. *Pointwise Mutual Information - PMI*) (Turney, 2002), фразе и ручно формираног скупа фраза везаних за електронске производе (нпр. „у камери“, „камера има“, „део камере“, итд.). Аутори су евалуирали свој приступ на ручно анотираном корпусу рецензија о електронским производима са *Amazon* сајта и саопштили просечан одзив од 0,77 и просечну прецизност од 0,94. На основу издвојених аспеката и синтаксичких зависности аутори су ручно формирали правила за издвајање фраза које носе сентимент, које су затим користили за идентификацију сентимент поларитета. За сентимент анализу аутори су постигли просечан одзив од 0,89 и просечну прецизност од 0,86.

За разлику од претходно описаних приступа, аутори (Thet *et al.*, 2010) су извршили фину гранулирану анализу како би одредио сентимент поларитет и интензитет сентимента у зависности од различитих аспеката филма (Филм, Директор, Глумац, Прича, Сцена и Музика). Фино гранулирана анализа подразумева анализу сегмента реченице мањег од целе реченице – ниво фраза и клауза (Секција 2.2.1.1). У циљу идентификације клауза у реченици, аутори су прво анализирали синтаксне зависности у реченици. Стабло зависности у реченици је формирано на основу Станфордовог *CoreNLP* алата (Manning *et al.*, 2014). Помоћу синтаксних зависности су реченице дељене на клаузе по знаковима интерпункције и везницима. Затим су клаузама додељивали један од предефинисаних аспеката и сентимент поларитета. Аспекти су идентификовани коришћењем ручно сакупљених речника термина везаних за аспекте филма. На пример, за аспект Прича аутори су сакупили синониме и термине везане за тај аспект попут сценарио, дијалог, романса, приповедач, завршница, заплет,

преокрет, драма итд. Затим су применили стемовање вршили поређење са терминима из речника. У случају подударања корена речи клаузе са термином из речника, датој клаузи је додељен одговарајући аспект. На сличан начин је свакој речи из клаузе додељен сентимент поларитет применом доменско специфичних лексикона и општих лексикона стентимент речи добијених из сентимент мреже речи *SentiWordNet*-а[18]. Тиме су свакој речи у клаузи доделили сентимент и на крају сумирали коначни сентимент поларитет клаузе. Аутори су своје истраживање евалуирали на 1.000 коментара филмова, које су ручно анотирали за одговарајуће аспекте и сентимент поларитете. За задатак идентификације аспеката су саопштене Ф-мере у опсегу од 0,94 до 0,98, док су за доделу сентимент поларитета саопштене просечне Ф-мере по аспектима у опсегу од 0,75 до 0,90.

Описани приступи показују да модели засновани на речницима и правилима постижу задовољавајуће резултате у *ABSA* области. Међутим, мана ових приступа је велика зависност од количине и квалитета анотираних ресурса. Такође, ови приступи имају велику зависност од домена и прилагођавање другом домену изискује велико време и труд доменских стручњака.

3.2 ABSA базирана на алгоритмима машинског учења

Бројне студије су истраживале потенцијалне предности употребе *ML* алгоритама у *ABSA*. Раније студије су се фокусирале на добро утврђене *ML* моделе обучених на ручно изграђеним језичким особинама (Alghunaim *et al.,* 2015; Jin *et al.,* 2009; Li *et al.,* 2010; Liu, 2015; Pang and Lee, 2008). У истраживању (Jin *et al.*, 2009) је *ABSA* задатак представљен као проблем секвенцијалног означавања применом скривених Марковљевих модела (енгл. *Hidden Markov Models* - *HMM*) са више језичких особина, као што су ознаке врсти речи и језичке особине из речника (нпр. синоними, антоними, речи сличног значења). Аутори су анализирали различите аспекте дигиталних камера из корпуса рецензија производа са Амазон сајта и додељивали им позитивни или негативни сентимент поларитет. Аспекте су прво груписали по ентитетима (компоненте, функционалности, карактеристике) и ручно означили речи и фразе у корпусу. Тако да су на пример, за ентитет карактеристике означавали речи боја, брзина, величина, тежина, јасноћа (слике дигиталне камере) итд. Затим су користећи речнике синонима и антонима проширили скуп аспеката на основу ког су аутоматски означавали

[18] Речник сентимент речи *SentiWordNet*: https://github.com/aesuli/sentiwordnet

(лабелирали) реченице у обучавајућем скупу података. Такође, ручно су означавали изразе у обучавајућем скупу који носе одређени сентимент (позитивне и негативне речи и фразе). Резултат анотације су реченице из обучавајућег скупа које су таговане на нивоу речи и фраза. Затим су обучавали *HMM* модел да на основу улазне секвенце речи пронађе одговарајућу секвенцу тагова. Сентимент поларитет је додељиван најближем препознатом аспекту у реченици на основу сентимент тагова добијених од *HMM* модела. Аутори су за задатак идентификације аспеката саопштили Ф-мере у опсегу од 0,75 до 0,83 и Ф-мере за задатак класификације сентимент поларитета у опсегу од 0,66 до 0,77.

У другом приступу, Li *et al.* (2010) су задатке *ABSA* такође моделовали као проблем секвенцијалног означавања али су користили модел условљених случајних поља (енгл. *Conditional Random Fields* - *CRF*) са различитим језичким особинама, као што су ознаке врсте речи и присуство синонима, антонима и сентимент речи. За одређивање синонима и антонима су користили су мрежу речи (*WordNet*), док су сентимент речи преузимали из мреже речи сентимента (*SentiWordNet*). Како би одредили дуже зависности (по питању броја речи) између аспеката и сентимент поларитета у реченици, аутори су користили синтаксичке зависности и коњукције између речи. Затим су реченице из обучавајућег скупа таговали поменутим језичким особинама и обучавали *CRF* модел да предвиди излазну секвенцу тагова. Приступ је евалуиран на корпусу рецензија о филмовима и производима. Укупна заједничка Ф-мера аспеката и сентимента је 0,77 за рецензије о филмовима и 0,79 за рецензије о производима. *CRF* модел је на корпусима рецензија филмова и производа надмашио тадашње моделе који су постизали најбоље перформансе у *ABSA* области (Hu and Liu, 2004; Jin *et al.*, 2009).

Веома важно такмичење које је допринело развоју *ABSA* је *SemEval*[19] такмичење на ком се од 2013. године врши надметање тимова за разне *NLP* задатке попут сентимент анализе, анализе хумора, ироније, сарказма и типа емоција за разне домене и језике. Аутори (Pontiki *et al.,* 2014, 2015, 2016) су за период од 2014. до 2016. године истакли да су се методологије најбоље рангираних тимова на *ABSA* задатку базирале на моделе машинског учења са богатим ручно развијеним језичким особинама (Hercig *et al.,* 2016; Kiritchenko *et al.,* 2014; Saias, 2015). На такмичењу 2017. године (Rosenthal *et al.*, 2017) тимови који су остварили најбоље резултате на задатку сентимент анализе су користили дубоке неуронске мреже. Тим (Cliche, 2017) који је остварио најбоље резултате је користио више варијанти *CNN* и *LSTM* модела, чије просечне перформансе су рангиране на првом месту. Међутим, дати задатак није подразумевао експлицитно

[19] Сајт *SemEval* такмичења за различите године: http://alt.qcri.org/semeval*GGGG*/ (уместо *GGGG* у линку унети годину)

издвајање аспеката него су они подразумевани и дати са подацима. У периоду од 2018. до 2021. године није било задатака из *ABSA* области, где је 2020. године једино био задатак сентимент анализе на подацима добијених мешањем језика (Patwa *et al.*, 2020). На основу датог се може закључити да *ABSA* задаци нису више у фокусу *SemEval* такмичења од 2018. године. Разлог може бити недостатак анотираних корпуса из домена и језика који се нису појављивали на такмичењима а за које постоји интересовање *NLP* заједнице.

Већина новијих истраживања на пољу сентимент анализе је усмерила пажњу на примену дубоких неуронских мрежа у *ABSA* (Dohaiha *et al.,* 2019). Један од првих приступа примене дубоког учења у *ABSA* је презентован у Wang and Liu (2015). У овом истраживању је анализиран СемЕвал'15 корпус (Pontiki *et al.*, 2015) од 550 рецензија о лаптоп рачунарима и ресторанима и вршена анотација одговарајућим аспектима и сентимент поларитетима на нивоу реченица. Аутори су обучавали потпуно повезану неуронску мрежу са два слоја како би идентификовали аспекте у реченици и онда применили *CNN* модел за доделу сентимент поларитета идентификованим аспектима. Свака реченица корпуса је репрезентована просечним векторима речи добијених *word2vec* алгоритмом обученим на корпусу Гугл вести. Улаз потпуно повезане вештачке неуронске мреже је просечни вектор речи на нивоу реченице, а излаз чине вероватноће за деветнаест аспеката. Предикција да реченица садржи одређени аспект је случај када је вероватноћа већа од прага који је одабран на валидационом скупу. Улаз *CNN* модела су векторске репрезентације реченице а излаз сентимент поларитет. Њихов приступ је евалуиран на СемЕвал'15 корпусу и резултати су поређени са *ML* моделом победничког тима СемЕвал'15 такмичења. Саопштили су Ф-меру од 0,51 за идентификацију аспеката и тачност од 0,78 за доделу сентимент поларитета. Аутори су за задатак идентификације аспеката надмашили перформансе модела победничког тима СемЕвал'15 такмичења за један проценат док исто толико подбацили за задатак доделе сентимент поларитета. Од тада је предложено више различитих приступа заснованих на *CNN* (Gu *et al.,* 2017; Ruder *et al.,* 2016; Wang and Liu, 2015; Wu *et al.,* 2016a; Xue and Li, 2018) и *RNN* моделима (Al-Smadi *et al.,* 2018; Ma *et al.,* 2018; Saeidi *et al.,* 2016; Yang *et al.*, 2018; Wang *et al.,* 2019). Одређени *RNN* приступи (Yang *et al.*, 2018; Wang *et al.,* 2019) су засновани на трансферу учења где су постојећи језички модели дообучени за одређени *NLP* задатак и домен.

Аутори (Yang *et al.*, 2018) су дообучили *ULMFiT* модел за *ABSA* задатак у домену финансија. Идентификацију аспеката и сентимента су вршили засебним моделима на нивоу реченице. Детекцију аспеката су вршили на два нивоа. Први ниво чине четири аспкета највеће апстракције, а то су Корпорација, Економија, Тржиште и Акције. Други ниво чини двадесет седам специфичнијих аспеката попут Састанци, Ризици, Финансије, Право, цена акције, и тако даље. Детекцију позитивног и негативног сентимента су представили као проблем регресије где је вршено предвиђање вредности у опсегу

$[-1, 1]$. Позитивне вредности ближе јединици су интерпретиране као више позитиван сентимент а обратно више негативан сентимент. Користили су *ULMFiT* модел који је претходно обучен на корпусу Википедије од 103 милиона речи. Аутори су дообучили модел на неанотираном корпусу из општег финансијског домена од тринаест милиона речи. Након генералног дообучавања модела су посебно додатно дообучили моделе за *ABSA* на анотираном корпусу од 1.174 реченица из вести и Твитер објава из домена финансија. Свака реченица из корпуса је садржала тачно по један аспект из првог и другог нивоа апстракције и један сентимент резултат. Модел аспеката су прво дообучили за први ниво аспеката а након тога трансфером знања формирали други модел који су дообучили за аспекте другог нивоа апстракције. Излаз модела аспеката су потпуно повезане вештачке неуронске мреже са истим бројем излаза колико и аспеката по нивоу апстракције. Модел сентимента су изменили тако што је уместо декодер дела генерално дообученог модела имплементирана потпуно повезана вештачка неуронска мрежа са бинарним излазом. Резултате *ABSA* су поредили са више модела засебно за задатке сентимент и аспектне анализе. Најбољу Ф-меру за задатак идентификације аспеката првог нивоа је постигао *ULMFiT* модел који је дообучен неанотираним корпусом из општег финансијског домена и износи 0,90. Док је *ULMFiT* модел који је додатно дообучен анотираним корпусом постигао најбољу Ф-меру од 0,75 за задатак идентификације аспеката другог нивоа. Најмању *MSE* грешку регресије од 0,08 је постигао *ULMFiT* модел који је додатно дообучен анотираним корпусом за задатак идентификације сентимент резултата. Постигнуте резултате су такође поредили са основним моделима машинског учења, моделом логистичке регресије за задатак идентификације аспеката и моделом линеарне регресије за задатак идентификације сентимент резултата. Модел логистичке регресије је остварио за два процента (0,02), односно за једанаест процената (0,11), мању Ф-меру од најбољег модела дубоког учења за идентификацију аспеката првог, односно другог, нивоа апстракције. Модел линеарне регресије је остварио за пет процената (0,05) већу *MSE* грешку регресије од најбољег модела дубоког учења.

У нешто другачијем приступу (Wang *et al.,* 2019), аутори су *ABSA* задатак вршили на основу хибридног модела дубоког учења. Предложени модел је сачињен од више аспект-сентимент капсула (Hinton *et al.*, 2011) којим су моделоване зависности између одређеног аспекта и сентимент поларитета. Аспект-сентимент капсула представља модул сачињен од групе вештачких неурона на основу којих се врше компликована интерна рачунања а резултати енкапсулирају у један вектор стања. На основу стања модула се рачунају вероватноће идентификације одређеног аспекта и сентимента. Предложени модел се обучава да на основу улазне секвенце предвиди парове аспект-сентимент који одговарају датој секвенци. На примеру реченице из домена ресторана „Особље није толико љубазно али укус хране то надокнади.“ се очекује излаз аспект-сентимент парова {(храна, позитиван), (услуга, негативан)}. Дакле, за сваки аспект се

формира једна капсула која предвиђа сентимент поларитет (позитиван, негативан и неутралан). Речи улазне секвенце су репрезентоване *Glove* моделом које су затим пропуштене кроз двосмерну *LSTM* мрежу, чије скривено стање се прослеђује свакој капсули. Затим се у капсули формира вектор стања на основу ког се израчунају вероватноће за дати аспект и сентимент поларитет. Све капсуле користе додатну заједничку двосмерну *LSTM* мрежу за међусобну комуникацију како би се спречило чување истих информација о секвенци у векторима стања капсула. Излази капсула су спојени у изланом модулу модела који даје предвиђене аспект-сентимент парове. Евалуацију модела су извршили на *SemEval'14* корпусу на домену ресторана. За задатак идентификације аспекта предложени модел је остварио просечну Ф-меру од 0,87, док је за задатак идентификације сентимента остварена просечну Ф-меру од 0,85.

Sun *et al.*, (2019) су *ABSA* задатак представили као проблем класификације пара реченице, као што је задатак аутоматског одговарања на питања. На основу анотација корпуса о аспектима и сентимент поларитету су формирали додатну реченицу, која је заједно са анотираном реченицом чинила пар реченица којим су дообучили *BERT* модел на задатку класификације две реченице. Циљ дообучавања *BERT* модела био је да на основу реченице (питање) предвиди реченицу (одговор) која носи *ABSA* информације. На пример, за један пар аспекта (генерални, цена, транзитна локација, сигурност) и сентимент поларитета (позитиван, негативан, ниједан) су формирали додатну реченицу попут „поларитет аспекта сигурност је позитиван". Аутори су обучавали додатне *BERT* моделе за задатак класификације једне реченице, где су за улазну реченицу једним моделом предвиђали аспекте а другим моделом предвиђали сентимент поларитет. Циљ аутора је упоређивање *BERT* модела са различитим задатком класификације у *ABSA* области. Евалуацију модела су извршили на *SentiHood* и *SemEval'14* корпусима и своје резултате поредили са *ML* моделом и више *RNN* модела без трансфера знања. На корпусу *SentiHood BERT* модел са задатком класификације пара реченица је остварио најбоље просечне резултате, и то Ф-меру од 0,88 за аспект и тачност 0,93 за сентимент анализу. Тиме су надмашили и додатни *BERT* модел са класификацијом једне реченице, и то за по седам процената за аспект и сентимент анализу. Нешто мању разлику су остварили у односу на остале моделе на корпусу *SemEval'14*. На основу резултата, аутори су закључили да је представљање *ABSA* задатак као проблем класификације пара реченица допринело бољим перформансама модела.

Постоји одређен број скоријих *ABSA* студија (Bhatnagar *et al.,* 2018; Gupta *et al.,* 2019; Mowlaei *et al.*, 2020; García-Díaz *et al.*, 2020) који показују да класични модели машинског учења и приступи базирани на природној обради језика, правилима и речницима могу и даље бити веома ефективни. Међутим, перформансе датих приступа нису упоређене са актуелним моделима дубоког учења попут *BERT* модела, који се показао веома успешно у моделовању језика и разним *NLP* задацима. Стога, при развоју

ABSA модела за различите домене и језике, потребно је прво испитати границе перформанси класичних модела машинског учења и модела заснованих на постојећим језичким ресурсима (нпр. речници, правила, онтологије) и упоредити са перформансама актуелних модела дубоког учења.

На основу досадашњег истраживања се може закључити да су модели дубоког учења са трансфером знања допринели бржем развоју *NLP* модела у *ABSA* области и постизање боље репрезентације језика него класични модели машинског учења. Такође, у *ABSA* области се нису још појавила истраживања која укључују новије вишејезичне моделе попут *XLM* и *XLM-RoBERTa* модела који могу допринети развоју *ABSA* модела и за друге мање познате језике.

3.3 ABSA у високом образовању

Већина истраживања у области *ABSA* у домену високог образовања је базирана на анализи на нивоу документа (Chauhan *et al.,* 2018; Valakunde and Patwardhan, 2013) и нивоу реченице (Shaikh and Doudpotta, 2019; Sindhu *et al.,* 2019). По најбољем сазнању аутора ове докторске дисертације, не постоје приступи који врше *ABSA* на нивоу сегмента реченице.

Аутори (Valakunde and Patwardhan, 2013) су применили *ABSA* на званичним студентским анкетама о перформансама факултета на нивоу документа. Као први корак, аутори су у свакој рецензији идентификовали предефинисане аспекте везане за наставно особље (знање, презентација, комуникација и регуларност одржане наставе). Аспекти су идентификовани на начин тако што је вршена претрага скупа ручно прикупљених кључних речи (нпр. знање, способност, стручност, итд.). Сентимент поларитет рецензије на нивоу документа је установљен применом *ML* класификатора обученом са језичким особинама заснованим на фреквенцији појављивања речи у документу. Аутори су за процес евалуације ручно анотирали корпус од 5.000 рецензија, прикупљених са универзитета аутора у Индији. Аутори су експериментисали са *NB* и *SVM* класификаторима и постигли најбољу тачност од 0,81 за сентимент поларитет са *SVM* класификатором. Перформансе идентификације аспеката нису саопштене засебно него су аспекти коришћени као тежински фактори приликом евалуације сентимента у документу.

У другом истраживању, аутори (Chauhan *et al.*, 2018) су извршили *ABSA* на нивоу документа студентских рецензија постављених на друштвене мреже. Студенти су коментарисали квалитет образованог система на универзитету аутора. Аутори су се

ослањали на именице и именичке фразе заједно са онтологијом концепата о високом образовању како би издвајали термине аспеката. Затим су на основу сентимент лексикона издвајали речи из документа које носе сентимент поларитет. Након тога су применом *Stanford CoreNLP* алата анализирали синтаксне зависности издвојених термина аспеката и сентимент речи, и на основу тога додељивали сентимент поларитет термину аспекта. Описани приступ је евалуиран на ручно анотираном корпусу од 1.000 рецензија са друштвених мрежа. Саопштена је Ф-мера од 0,80 за идентификацију аспеката и 0,72 за индентификацију сентимент поларитета.

У нешто скоријем истраживању, аутори (Shaikh and Doudpotta, 2019) су представили хибридни *ABSA* приступ заснован на моделима *ML* и правилима. *ML* модел је обучен коришћењем приступа вреће речи (*BOW* приступ) при класификацији реченица у један од два предефинисана аспекта (наставник или предмет). Након тога је примењен модел заснован на правилима како би се реченице додатно класификовале на финије категорије аспеката (нпр. понашање, знање, искуство, итд.). Затим су на основу правила и *Stanford CoreNLP* алата издвајали придеве и одређивали сентимент поларитет коришћењем *SentiWordNet*-а. На основу тога су додели сентимент поларитет сваком идентификованом аспекту. Аутори су ручно анотирали корпус од 10.000 званичних студентских анкета на пакистанском језику како би евалуирали њихову методологију. За задатак идентификације аспеката је саопштена просечна прецизност од 0,83 и просечан одзив од 0,80. Док је за задатак сентимент анализе саопштена просечна тачност од 0,90.

Аутори (Sindhu *et al.*, 2019) су у свом истраживању описали *ABSA* на нивоу реченице базирано на дубоком учењу. Аутори су користили *LSTM* моделе како би идентификовали аспекте (педагогија у настави, понашање, знање, процена знања, искуство и уопштено) и сентимент поларитет. Применом *Skip-gram* верзије *Word2Vec* (Секција 2.2.1.3.1) модела су формиране посебне векторске репрезентације речи прилагођене за домен високог образовања, које су коришћене при обучавању модела. Аутори су евалуирали предложени модел на ручно анотираном корпусу од 5.000 анкета са *Sukkur IBA* универзитета као и на добро познатом *SemEval'14* корпусу за домен који није из високог образовања. Саопштена је просечна Ф-мера од 0,85 за идентификацију аспеката и просечна Ф-мера од 0,86 за идентификацију сентимент поларитета на корпусу аутора. Евалуацијом модела на *SemEval'14* корпусу рецензија из домена ресторана постигнута просечна Ф-мера од 0,82 за задатак идентификације аспеката и просечна тачност од 0,85 за задатак идентификације сентимент поларитета. Поред тога, аутори су поредили перформансе предложеног модела са перформансама других класичних модела машинског учења. Предложени модел дубоког учења је надмашио перформансе *SVM* модела (са 13 процената већом тачношћу за сентимент), *NB* модела

(са 4 процената већом тачношћу за сентимент и аспекте) и *NB* модела са лексиконима (са 13 процеанта већом тачношћу за сентимент и 8 процената за аспекте).

Као што се може видети, већина ранијих приступа у *ABSA* у високом образовању се заснива на хибридним (*ML*, речници и правила) методологијама, док недавне студије истражују примену дубоког учења. Међутим, ни један од описаних приступа не спроводе *ABSA* на детаљнијем и софистицираније нивоу анализе (сегменту реченице), поготову за српски језик. У том смислу, истраживање приказано у овом раду представља нов допринос *ABSA* пољу у домену високог образовања.

3.4 Сентимент анализа у српском језику

Сентимент анализа у српском језику представља релативно нову област истраживања у погледу доступних научних радова. Модели за аутоматску анализу сентимента на српском језику почели су да се развијају од 2012 године (Milošević, 2012). Досадашњи приступи (Milošević, 2012; Mladenović *et al.*, 2016; Batanović and Nikolić, 2017; Grljević, 2016; Grljević *et al.*, 2020; Kovačević *et al.*, 2020) се заснивају само на анализи сентимента на нивоу документа и реченице. Међутим, за прецизније извршавање *ABSA* задатака је потребно анализирати и аспекте на детаљнијем нивоу анализе текста од нивоа реченица – нивоу сегмента реченице (фраза и клауза).

Међу првим истраживањима сентимента на српском језику је мастер рад аутора (Milošević, 2012). Аутор је у вршио класификацију реченица на српском језику према сентимент поларитету (позитиван, негативан) уз помоћ *NB* модела. Аутор је у оквиру свог истраживања развио хибридни модел за свођење речи српског језика на њен корен (стемовање), уклањањем суфикса речи (Milošević, 2012). Евалуација модела је извршена на корпусу реченица општег домена и остварена је прецизност од 0,95.

Аутори (Mladenović *et al.*, 2016; Mladenović, 2016) су анализирали сентимент применом *ML* модела користећи језичке особине сентимент лексикона и *WordNet*[6] речника за српски језик (Krstev *et al.*, 2004). Евалуација модела је извршена на два корпуса из домена вести и једним корпусом из домена филмова. Најбоље резултате класификације сентимента на нивоу документа аутори су остварили коришћењем карактеристика фраза односно униграма и биграма, допуњеним језичким особинама сентимент речника. Саопштена је тачност модела од 0,78 за рецензије филмова и 0,79 за рецензије вести.

Аутори (Batanović and Nikolić, 2017) је анализирао утицај морфолошке нормализације текста на класификацију докумената на српском језику према

поларитету сентимента. Коришћен је *BOW* приступ где је документ рецензије филма моделован као скуп речи (*n-gram*). Евалуација је извршена за три модела: *MNB*, *SVM* и хибридни модел сачињен од претходна два модела. Најбоља тачност класификације сентимента на позитиван и негативна поларитет од 0,85 је остварена коришћењем хибридног модела и *NLP* алата за свођење на корен речи аутора (Milošević, 2012). Аутори су свођењем речи на њен корен смањили број морфолошких облика речи (нормализација) у врећи речи и тиме допринели побољшању тачности од 1,24 за позитиван и негативан поларитет.

Ауторка докторске дисертације (Grljević, 2016) је анализирала сентимент у студентским рецензијама наставног особља на сајтовима друштвених мрежа и медија, у циљу унапређења пословања високошколских установа. Рецензије су преузете са сајта оцени професора и ручно анотиране на нивоу реченице са аспектом рецензирања, сентимент оријентацијом, интензитетом исказаног сентимента, речима на основу којих је утврђен сентимент, негацијом и њеним опсегом важења. Овако припремљен и анотиран корпус је додатно обрађен и представљен врећом речи (*n-gram*) укључујући изразе од једне до четири узастопне речи. Сентимент анализа коментара је извршена на нивоу реченица применом алгоритама надгледаног учења *MNB*, *SVM* и *k-NN*. Поред *ML* модела је формиран и модел заснован на сентимент речницима. Коришћено је пет речника изведених на основу анотираних сентимент речи, кључних речи негације, као и на основу запажања изведених током анотације. За сваку реченицу су примењени речници позитивног и негативног сентимента, где је поларитет мењан речником негације. На пример, уколико се над једном реченицом примени два пута речник позитивног сентимента, резултат је позитиван са интензитетом два. У случају присуства негације, укупан резултат се мења на негативан са интензитетом два. Додатно су примењени речници појачавања и смањивања интензитета одређеног поларитета сентимента. Применом речника појачавања интензитета сентимента су укупни резултат множили са два, док су речници смањивања интензитета сентимента укупни резултат делили са два. Задатак класификације сентимента реализован је на два нивоа грануларности: на нивоу документа и на нивоу реченице. Највећа тачност сентимент анализе је постигнута *SVM* моделом од 0,85 на нивоу документа и 0,81 на нивоу реченице. Ауторка је анализирала могућност употребе развијених модела идентификације сентимента у пословању високошколских установа, на примеру Економског факултета у Суботици. Закључак истраживања је да овакви модели омогућавају поређење онлајн репутације установе са онлајн репутацијом конкуренције, као и идентификовање извора задовољства и незадовољства студената.

На основу претходних истраживања се може закључити да су *ABSA* системи зависни од домена и језика. Најбоље резултате дају системи засновани на *ML* моделима и употреби *NLP* алата и ресурса и ручно изграђених правила, речника и лексикона. За

развијање *ML* модела је неопходан анотирани корпус чија изградња је временски захтевна. Модели базирани на дубоком учењу захтевају постојање већег корпуса и *NLP* алата и ресурса за језик који се анализира. *ABSA* системи за српски језик су заступљени само за задатке анализе сентимента, не и аспеката као носиоца сентимента. Тачније, ауторка (Grljević, 2016) је поред сентимент анализе вршила анотирање корпуса аспектима студирања и статистички приказ, али није разматрала развој аутоматизованих метода за класификацију. Такође, досадашње анализе сентимента су вршена на целим документом (Mladenović *et al.*, 2016; Batanović and Nikolić, 2017) или реченици (Grljević, 2016). У циљу прецизније и тачније анализе сентимента је потребна анализа на финијем нивоу грануларности као што је сегмент реченице (фраза, клауза). Из угла аспектно базиране сентимент анализе не постоји систем за аутоматску анализу мишљења за српски језик.

4 Корпус златног стандарда

Корпус који је коришћен за обучавање и евалуацију модела ове докторске дисертације представља колекцију коментара студената о наставном особљу и студијским програмима Факултета техничких наука, Универзитета у Новом Саду (скраћено ФТН). Додатно, модел је евалуиран корпусом кога чини колекција јавно доступних коментара наставног особља разних факултета у Србији са веб сајта "Оцени професора" (Kovačević *et al.*, 2020).

Оба корпуса су из домена високог образовања који садрже коментаре на српском језику и сакупљени су ручно из два извора података. Први корпус (К1) представља колекцију рецензија добијених из званичних студентских анкета спроведених на ФТН-у током периода од 2011. до 2016. године. Анкете су прибављене уз сагласност управе ФТН-а и анонимизоване, како би се сачувала приватност података[20]. Анкетирање студената је обавезно у образованим установама и спроводи се сваке године према важећем закону о образовању. Анкете су попуњаване ручно а потом су ручно преписане у базу података од стране административног особља факултета. Процес прикупљања података и формирања корпуса К1 се заснивао на издвајању рецензија из анкета и додељивању доступних мета података попут идентификатора предмета и наставника, датума анкетирања и тако даље. На тај начин је сачувана веза са структуираним подацима из анкета који ће бити неопходни за анализу резултата предложеног система докторске дисертације у Секцији 7. Корпус К1 садржи 2.472 коментара студената, тачније 2.798 реченица или 21.028 речи. Дужина коментара варира од 2 до 80 речи, а нешто краћи коментари (3 до 15 речи) чине 74 процената корпуса.

Други корпус (К2) представља колекцију од 3.863 јавно доступних коментара са интернета који су прикупљени за период од 2012. до 2016. године. Дужина коментара варира од 1 до 180 речи, а коментари средње дужине (6 до 30 речи) чине 53 процената

[20] Закон о заштити података о личности, „Службени гласник РС", бр. 97/2008, 104/2009, 68/2012 – Одлука Уставног суда, 107/2012

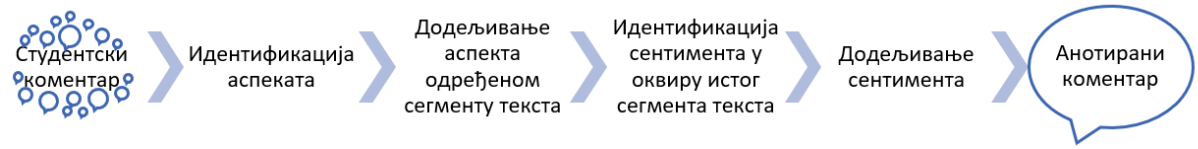
корпуса. Коментари у корпусу К2 садрже 6.896 реченица. За разлику од корпуса К1, рецензије корпуса К2 су постављане на Интернет слободном вољом корисника веб сајта (под претпоставком студената) за одабране професоре било ког факултета у Републици Србији. Детаљнија анализа корпуса К2 се може пронаћи у истраживању аутора (Grljević, 2016).

У наставку је описан процес анотације корпуса атрибутима аспекта и сентимент поларитета. Након тога је описана анлаиза поузданости анотатора који су учествовали у прецосу анотације корпуса. На самом крају поглавља се налази статистичка анализа корпуса.

4.1 *Анотација корпуса К1*

Анотација корпуса К1 је обављена по истој процедури и шеми анотације који су коришћени у истраживању (Grljević, 2016) за корпус К2. У анотацији корпуса К1 су учестовала два анотатора која су образована на Универзитету у Новом Саду и тиме упознати са терминима високог образовања. Анотатори су првобитно обучени за анотацију на насумично одабраних десет процената корпуса (240 рецензија) по упуству из истраживања (Grljević, 2016). Анотатори су додатно упућени да се приликом анотирања сегмената реченица обрати пажња на границе клауза и фраза (дефинисани у наставку). Затим је извршено независно анотирање од стране оба анотатора и израчунато слагање анотатора. На основу великог процента сагласности анотатора процес анотације је настављен. Цео процес обучавања и анотације је вршен у просторијама ФТН-а и укупно је трајао три недеље.

Корпус К1 је анотиран применом MAE алату[21] по анотационој процедури (Слика 53) представљених у истраживању (Grljević, 2016).

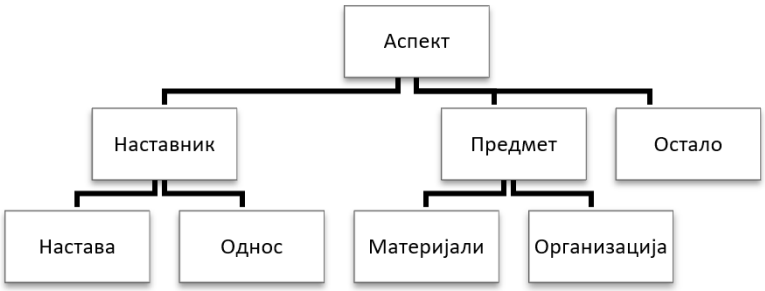


Слика 53 *Процедура анотације корпуса*

[21] MAE алат за анотацију корпуса: https://github.com/amber-stubbs/mae-annotation

Главни циљ ове шеме анотације је био да се обухвати што је више могуће информација из коментара студената написаних у формату слободног текста. Корпуси су анотирани атрибутима аспекта и сентимента на нивоу сегмента текста. За сваки текст, који је предмет коментара, анотатори су морали да доделе тачно један аспект и један сентимент. Анотирани сегменти текста су у оквирима реченице. У случају просте реченице анотирани сегмент текста представља целу реченицу. Док се у случају сложене реченице анотира више сегмената текста који представљају сегменте реченице. У полгеду граматичке структуре, анотирани сегменти текста могу бити клаузе и фразе. Клаузе садрже субјекат и предикат, и могу бити независне и зависне клаузе. Независне клаузе су еквиваленти простим реченицама, док зависне улазе у састав сложених реченица. Док су фразе мање целине од клауза и не садрже везу између субјекта и предиката. За корпус К1 је начињено укупно 3.403 анотација. Више детаља о атрибутима анотације, поступку означавања и тумачењу резултата, дати су у наставку.

Аспекте студирања у високом образовању могу чинити особе, ствари или понашање које чини објекат тј. носиоца мишљења. На основу шеме из претходног истраживања аутора (Grljević, 2016) је дефинисано шест аспеката студирања: *Наставник*, *Настава*, *Однос*, *Предмет*, *Материјали* и *Организација*. *Наставник* на факултету представља особу која учествује односно одржава наставу на одређеном предмету. Под аспектом *Наставник* се убрајају све запослене особе које учествују у настави без обзира на научно звање попут професора, асистента, сарадника у настави, лаборанта и демонстратора. Сви облици наставе попут предавања, аудиторних вежби, лабораторијских вежби, испита и праксе су обухваћени аспектом *Настава*. Понашање и опхођење наставника према студентима током наставе и консултација је представљено аспектом *Однос*. Сваки предмет на факултету има курикулум којим је описана организација предмета и предложена литература. Стога аспект *Организација* подразумева начин организовања свих облика наставе и оцењивања на одређеном предмету. Сва литература која се користи на предмету попут књига, скрипти и других докумената коришћених у настави су обједињени аспектом *Материјали*. Одређени аспекти представљају апстракцију других аспеката, те је стога дефинисана хијерархијска шема анотације за атрибут анотације *Аспект* и приказана на (Слика 54).



Слика 54 *Хијерархијска шема антоације аспекта студирања*

Према предложеној шеми анотатори су упућени да прате приступ одоздо према горе за анотацију аспеката. Аспекти вишег нивоа (*Наставник*, *Предмет*) треба користити само у случајевима када се аспекти нижег нивоа (*Настава*, *Однос*, *Материјали*, *Организација*) не могу идентификовати. Аспект *Остало* је додељен само у случајевима када се не могу идентификовати аспекти студирања, на пример за коментаре као што су „Спава ми се“, „Исцрпљен сам“, „Све је ОК“, „Све најбоље“ и тако даље. Сваком сегменту текста коме је додељен аспект мора бити анотиран и са атрибутом сентимента. Атрибут сентимента може имати једну од три унапред дефинисане вредности: позитиван, негативан или неутралан. Пример резултата анотације једног студентског коментара применом MAE алата је дат на Слика 55.

```
<?xml version="1.0" encoding="UTF-8" ?>

<AnotiranjeKomentara>
<TEXT><![CDATA[Predmet je bitan ali loše koncipiran.
Preobimno gradivo za jedan semestar
Profesor preobimno objašnjava i predaje.
]]></TEXT>
<TAGS>
<ASPEKT id="A1" spans="18~35" text="Predmet je bitan "
type="predmet" sentiment="pozitivan" />
<ASPEKT id="A2" spans="34~55" text=" ali loše koncipiran."
type="organizacija" sentiment="negativan" />
<ASPEKT id="A3" spans="56~91" text="Preobimno gradivo za jedan semestar"
type="organizacija" sentiment="negativan" />
<ASPEKT id="A4" spans="92~132" text="Profesor preobimno objašnjava i predaje."
type="nastava" sentiment="negativan" />
</TAGS>
</AnotiranjeKomentara>
```

Слика 55 *Пример резултата анотације једног студентског коментара*

У случају корпуса К2 је поред сентимента и аспеката анотирано интензитет сентимента, сентимент израз, и негација. Сентимент изразе чини једна или више речи које носе позитиван или негативан сентимент. У истраживању (Grljević, 2016) је саопштено да је за корпус К2 направљено 8.101 анотација.

4.2 Анализа поузданости анотатора

Како би се проценило колико добро је дефинисан задатак анотације користи се мера међусобне сагласности анотатора. Уколико је мера сагласности велика, то указује да је задатак анотације добро дефинисан и да анотатори могу наставити процес анотације корпуса. Међутим, висока мера сагласности не значи нужно да су анотације тачне него да анотатори тумаче упутства анотације на исти начин (Pustejovsky and Stubbs, 2012). Тиме је установљена конзистентност и поузданост анотатора чије анотације се могу посматрати подједнако на нивоу анотираног корпуса. Мера сагласности анотатора се обично дефинише статистичком мером *Kappa* статистиком (Smeeton, 1985). За поређење сагласности два анотатора се најчешће користи *Cohen's Kappa* мера (Cohen, 1960; Sim and Wright, 2005), док за поређење сагласности више од два анотатора се користи *Fleiss Kappa* мера (Fleiss, 1971; Fleiss *et al.*, 2003). Међутим, анотатори у *NLP* области по сопственом нахођењу означавају секвенце текста те се стога може разликовати број анотација и анотиране секвенце текста. У том случају је по речима аутора (Ide and Pustejovsky, 2017) адекватније применити *agr* меру сагласности (Wiebe *et al.*, 2005).

Приликом анотирања корпуса К1, анотатори су начинили преко 90 процената анотација над истом секвенцом текста. Стога ће за дати проценат анотација у наставку бити израчуната *Cohen's Kappa* мера сагласности анотатора, која одговара анотираном корпусу од стране два анотатора. Додатно ће бити израчуната *agr* мера сагласности за све анотације, узимајући у обзир и оне анотације чије се анотиране секвенце текста разликују.

Слагање анотатора корпуса К1 је рачунато засебно за анотацију аспеката и сентимент поларитета ради бољег увида у извор неслагања. Аутор (Cohen, 1960) је предложио мерење сагласности два анотатора које узима у обзир вероватноћу несигурности анотатора при додељивању одређене лабеле. Како би се израчунао коефицијент *Cohen's Kappa* мере сагласности, прво је потребно организовати резултате два анотатора у унакрсној табели у матричном облику (Табела VIII). У зависности од броја класа који анотатори додељују за одређени задатак, димензија матрице (осенчане у табели) се мења. За случај овог истраживања димензија квадратне матрице је три за задатак анотације сентимента а седам за случај задатка анотације аспеката. Више детаља је описано у наставку.

Табела VIII *Унакрсна табела резултата два анотатора*

Анотатор 2 / Анотатор 1	Класа 1	Класа 2	…	Класа К	Укупно
Класа 1	b_{11}	b_{12}	…	b_{1k}	n_{11}
Класа 2	b_{21}	b_{22}	…	b_{2k}	n_{21}
…	…	…	…	…	…
Класа К	b_{k1}	b_{k2}	…	b_{kk}	n_{k1}
Укупно	n_{12}	n_{22}	…	n_{k2}	N

Вредности матрице у табели означавају број анотација за одређену класу добијених на следећи начин. Сагласности анотатора у вези доделе одређене класе биће смештени у једној од дијагоналних ћелија матрице. На пример, ако су се анотатори за одређени узорак из анотационог корпуса сложили да припада првој класи, онда се вредност b_{11} ћелије матрице увећава за један. Несагласност анотатора се смешта у једној од преосталих ћелија матрице без главне дијагонале. На пример, уколико је први анотатор одређени узорак из анотационог корпуса означио првом класом а други анотатор класом К, онда се вредност b_{1k} ћелије матрице увећава за један. Насупрот томе, уколико је први анотатор означио класом К а други анотатор првом класом, онда се увећава вредност b_{k1} ћелије матрице.

Након тога се израчунају суме редова и колона класа посматране матрице и резултат упише у одговарајући ред, односно колону, табеле (Табела VIII). На пример, сума првог реда матрице (ред „Класа 1“) првог анотатора је означена са n_{11}, док је сума прве колоне матрице (колона „Класа 1“) другог анотатора означена са n_{12}. Укупан број анотација учињених од стране сваког анотатора (означеног са N) одговара укупној суми сума редова ($\sum_{k=1}^{K} n_{k1}$) и укупној суми сума колона ($\sum_{k=1}^{K} n_{k2}$) табеле.

Коефицијент *Cohen's Kappa* мере сагласности тј. корелације два анотатора (означено са κ) се рачуна по следећој формули:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

Где p_o представља пропорцију узорака где су се анотатори сложили и p_e очекивану пропорцију узорака где су се анотатори случајно сагласили. Вредности коефицијента *Cohen's Kappa* мере корелације κ се крећу у опсегу од -1 до 1, при чему

вредност 0 представља сагласност која се може очекивати насумичним одабиром тј. вероватноћу случајног одабира, а 1 преставља потпуно слагање анотатора. Вредности од 0 до -1 означавају да не постоји ефективна сагласност између два анотатора тј. да је сагласност лошија од анотације насумичним одабиром лабеле или класе.

Пропорција слагања анотатора p_o се добија сумирањем вредности ћелија главне дијагонале матрице (Табела VIII) и дељењем укупним бројем анотација по анотатору. Док се пропорција случајне сагласности анотатора p_e рачуна као сума производа броја анотација оба анотатора за сваку класу засебно, подељених са укупним бројем анотација по анотатору. Изведене формуле аутора (Cohen, 1960) су дате у наставку:

$$p_o = \frac{1}{N}\sum_{k=1}^{K} b_{kk}$$

$$p_e = \frac{1}{N^2}\sum_{k=1}^{K} n_{k1} \times n_{k2}$$

Где *N* представља укупан број анотација по анотатору, *K* укупан број анотираних класа, b вредности матрице унакрсне табеле, n_{k1} и n_{k2} суме анотација класе *k* првог и другог анотатора (описаних у поступку формирања Табела VIII). Додатно, након израчунатог *Cohen's Kappa* коефицијента се рачуна његова стандардна грешка (енгл. *Standard Error - SE*) на следећи начин:

$$SE_k = \sqrt{\frac{p_o(1-p_o)}{N(1-p_e)^2}}$$

На основу описаног поступка формирања унакрсне табеле резултата анотатора (Табела VIII), резултати анотирања извршеног у овом истраживању су формирани у унакрсним табелама. У наставку је наведена унакрсна табела резултата анотације новатора аспеката студирања (Табела IX) описаних у Секцији 4.1.

Табела IX *Унакрсна табела резултата анотатора за аспекте*

Анотатор 2 / Анотатор 1	Наставник	Настава	Однос	Предмет	Материјали	Организација	Остало	Укупно
Наставник	44	10	3	1	0	0	0	58
Настава	7	43	0	1	0	11	1	63
Однос	2	0	9	0	0	0	0	11
Предмет	0	0	0	7	0	2	0	9
Материјали	0	1	0	0	12	1	0	14
Организација	0	10	0	2	1	87	0	100
Остало	1	2	0	1	0	2	2	8
Укупно	54	66	12	12	13	103	3	263

У датој табели се може видети да углавном постоји велика сагласност анотатора за све анотиране аспекте. Мања одступања сагласности се може уочити код аспеката *Наставник*, *Настава* и *Остало*. У наставку је наведена унакрсна табела резултата сентимент поларитета (Табела X) описаног у Секцији 4.1.

Табела X *Унакрсна табела резултата анотатора за сентимент поларитет*

Анотатор 2 / Анотатор 1	Позитиван	Негативан	Неутралан	Укупно
Позитиван	55	0	0	55
Негативан	2	201	3	206
Неутралан	0	1	1	2
Укупно	57	202	4	263

У датој табели се може уочити да постоји велика сагласност анотатора за сентимент поларитет. На основу формираних унакрсних табела резултата аспеката студирања (Табела IX) и анотације сентимент поларитета (Табела X) су израчунати коефицијенти κ *Cohen's Kappa* мере сагласности два анотатора и стандардне грешке *Kappa*-е (SE_k). Резултати *Cohen's Kappa* мере сагласности су наведени у Табела XI.

Табела XI *Резултати мере сагласности анотатора за сентимент поларитет и аспекте*

Анотација	f_o	f_e	κ	SE_k
Сентимент поларитет	0.98	0.65	0.94	0.02
Аспекти	0.78	0.26	0.70	0.03

Резултати *Cohen's Kappa* мере сагласности се могу тумачити на следећи начин. Коефицијент κ за задатак анотације сентимент поларитета од 0,94 означава да би два анотатора, за насумично одабрани сегмент реченице, били у сагласности у преко 94 процената случајева. На исти начин за задатак анотације аспеката студирања, у преко 70 процената случајева оба анотатора би произвољни сегмент реченице анотирали на исти начин. Мала вредност стандардне грешке *Kappa*-е (SE_k) указује на независност анотатора приликом анотације. Међутим, добијене вредности *Cohen's Kappa* коефицијента не одређују да ли је постигнута довољна сагласност анотатора и стога је потребно интерпретирати их.

Међу првим смерницама за интерпретацију резултата *Kappa* статистике су дали аутори (Landis and Koch, 1977) на следећи начин: вредности κ преко 0,81 показују скоро савршену сагласност, од 0,61 до 0,80 као јаку сагласност, од 0,41 до 0,60 као умерену, од 0,21 до 0,40 као слабу сагласност, од 0,00 до 0,20 као минималну, а мање од 0,00 као непостојање сагласности (неслагање). Према наведеној скали κ вредности се може закључити да је приликом анотације корпуса К1 постигнута скоро савршена сагласност за задатак анотације сентимент поларитета (*Cohen's Kappa* коефицијент од 0,94) и јака сагласност за задатак анотације аспеката студирања (*Cohen's Kappa* коефицијент од 0,70). Нешто новије смернице интерпретације резултата *Kappa* статистике дао је аутор (Krippendorff, 2018). За разлику од првобитног предлога аутора (Landis and Koch, 1977), скала κ вредности коју је предложио аутор (Krippendorff, 2018) се састоји и следећа три опсега: вредности преко 0,81 показују одличну сагласност, од 0,65 до 0,80 као задовољавајуће добру сагласност, а мање 0,65 као лошу сагласност анотатора. Према наведеној скали сагласности анотатора постигнута је одлична сагласност анотатора приликом анотирања сентимента (*Cohen's Kappa* коефицијент од 0,94) и задовољавајуће добра сагласност анотатора приликом анотирања аспеката (*Cohen's Kappa* коефицијент од 0,70) за корпус К1. На основу добијених резултата сагласности анотатора *Kappa* статистике је настављена анотација а након тога је корпус К1 примењен као корпус златног стандарда у сврху овог истраживања.

Уколико се анализу поузданости анотатора укључе и анотације чије се анотиране секвенце текста разликују, онда се у том случају примењује *agr* мера сагласности. Ако се скуп анотација анотатора $A1$ представи са $AN1$ а скуп анотација анотатора $A2$ са $AN2$, тада се рачунају две мере преклапања анотација на следећи начин:

$$agr(A1||A2) = \frac{|AN1 \cap AN2|}{|AN1|}$$

$$agr(A2||A1) = \frac{|AN2 \cap AN1|}{|AN2|}$$

Где $agr(A1||A2)$ мери слагање анотација два анотатора $A1$ и $A2$ када се за златни стандард узме $AN1$ скуп анотација. Док $agr(A2||A1)$ мери слагање анотација два анотатора када се за златни стандард узме $AN2$ скуп анотација. По начину рачунања мера преклапања, $agr(A1||A2)$ мера представља одзив а $agr(A2||A1)$ мера прецизност, стога се мера слагања анотатора добија рачунањем Ф-мере на начин описан у Секцији 2.4.1:

$$\text{Ф-мера} = \frac{2 * \text{прецизност} * \text{одзив}}{\text{прецизност} + \text{одзив}}$$

На основу броја анотација по аспектима и броју анотација где се подударају анотиране секвенце текста су израчунате мере сагласности анотатора и представљене у Табела XII. На основу резултата се може закључити да је највиша сагласност анотатора постигнута за аспекте *Наставник*, *Материјали* и *Организација*, док је најмања за аспект *Настава* и *Остало*.

Табела XII *agr мера сагласности анотатора за аспекте*

	Наставник	Настава	Однос	Предмет	Материјали	Организација	Остало
Ф-мера	0.75	0.60	0.69	0.56	0.73	0.79	0.31

На сличан начин су за сентимент поларитет израчунате мере сагласности анотатора и приказане у Табела XIII. На основу резулта се може видети да су анотатори имали већу сагласност за негативан него за позитиван сентимент.

Табела XIII *agr мера сагласности анотатора за сентимент*

	Позитиван	Негативан	Неутралан
Ф-мера	0.86	0.92	0.33

Вредности *agr* мере сагласности су ниже у односу на меру сагласности *Kappa* статистике јер ипак постоји десет процената неслагања анотатора. У истраживању

анотације корпуса у области сентимент анализе, аутори (Wiebe *et al.*, 2005) су за *agr* меру сагласности постигли Ф-мере у опсегу од 0,59 до 0,81 за задатак анотације субјективних елемената реченице. Док су аутори (Barnes et al., 2018) вршили анотацију два корпуса у *ABSA* области постигли Ф-мере у опсегу од 0,12 до 0,77. За задатак анотације аспеката аутори (Barnes et al., 2018) су остварили просечне Ф-мере у опсегу од 0,12 до 0,26, док су за задатак анотације сентимент израза остварили просечне Ф-мере у опсегу од 0,71 до 0,72. Аутори су мале просечне Ф-мере за задатак анотације аспеката образложили чињеницом да су већину корпуса чинили имплицитни изрази аспеката. Узимајући у обзир резултате *agr* мере сагласности поменутих истраживања се дошло до закључка да постоји велика сагласност анотатора корпуса К1 за задатак анотације сентимента док за задатак анотације аспеката постоји средња сагласност анотатора због аспеката *Настава* и *Остало*.

На основу података из истраживања (Grljević, 2016), корпус К2 је анотиран од стране четири анотатора, и стога је примењена *Fleiss Kappa* мера за одређивање слагања анотатора. На основу предложене скале сагласности анотатора од стране аутора (Krippendorff, 2018) постигнута је одлична сагласност анотатора приликом анотирања сентимента (*Fleiss Kappa* коефицијент од 0,96) и задовољавајућа сагласност анотатора приликом анотирања аспеката (*Fleiss Kappa* коефицијент од 0,79). Више детаља о анализи сагласности анотатора за корпус К2 се може наћи у истраживању (Grljević, 2016).

4.3 Статистичка анализа корпуса

Табела XIV и Табела XV приказују детаљније статистике за анотационе атрибуте аспекта и сентимент поларитета. Ради упоредне статистичке анализе корпуса К1 и К2, у обе табеле су додати статистички подаци из претходног истраживања аутора (Grljević, 2016).

На Табела XIV треба приметити да је 56,2 процената (1.836 анотација) корпуса К1 подједнако чине аспекти *Настава* и *Организација*, док је 58,0 процената (4.669 анотација) корпуса К2 анотирано класом аспекта *Наставник*. Најмање анотација првог корпуса је направљено за аспект *Остало* (3 процента), док за други корпус класе *Организација* (1 проценат), *Предмет* (2 процента), *Остало* (2 процента) и *Материјали* (3 процента) садрже најмање анотација. Ова диспропорција класа представља последицу различитог типа анкета два корпуса. Рецензије корпуса К2 су преузете са веб сајта „оцени професора“ стога је и очекивано да већину анотација чини аспект Наставник. Међутим, у званичним анкетама ФТН-а, студенти су на основу постављених питања подстакнути да коментаришу све аспекте образовања према њиховом искуству.

Табела XIV *Статистика корпуса златног стандарда за аспекате*

	Број анотација у односу на проценат свих анотација	
Аспект	**К1**	**К2**
Наставник	638 (19,0%)	4.669 (58,0%)
Настава	925 (27,5%)	1.186 (15,0%)
Однос	193 (6,0%)	1.538 (19,0%)
Предмет	204 (6,0%)	176 (2,0%)
Материјали	362 (11,0%)	280 (3,0%)
Организација	925 (27,5%)	108 (1,0%)
Остало	102 (3,0%)	144 (2,0%)

У Табела XV се може приметити асиметричност већине анотација сентимента два корпуса – 73,3 процената анотација корпуса К1 чине негативан сентимент, док је 61,0 процената анотација корпуса К2 позитивног сентимента. Разлог ове различите расподеле сентимента у корпусима се може повезати са окружењем у којем су настале рецензије. Студенти нису у обавези да користе веб страницу „Оцени професора“ и стога могу по својој слободној вољи и жељи одабрати наставника кога желе да коментаришу. Самим тим и начин изражавања током писања рецензије се може разликовати у поређењу са званичним и обавезним анкетама спроведеним на факултету. Анкетни процес на факултету се спроводи за сваки предмет засебно током семестра у ком су студенти похађали исти. Најчешће студенти попуњавају анкете током предавања уз присуство предметног наставника што се може узети као фактор утицаја на сам начин писања рецензија. Број неутралних анотација указује на високо субјективан текст у оба корпуса.

Табела XV *Статистика корпуса златног стандарда за сентимент*

	Број анотација у односу на проценат свих анотација	
Сентимент	**К1**	**К2**
Позитиван	901 (25,7%)	4.941 (61,0%)
Негативан	2.448 (72,3%)	3.051 (38,0%)
Неутралан	54 (2,0%)	109 (1,0%)

Распрострањеност позитивног поларитета у корпусу К2 је аутор (Grljević, 2016) објаснио феноменом *Pollyanna* (Boucher and Osgood, 1969). Феномен *Pollyanna* је повезан са изражавањем емоција у рецензијама на Интернету (Taboada, 2016), где су аутори истраживали људску склоност ка прецизнијем памћењу позитивнијег искуства у односу на непријатна искуства. Аутори су дошли до закључка да су рецензије на Интернету за различите домене претежно позитивне.

5 Имплементација система за аутоматску анализу мишљења

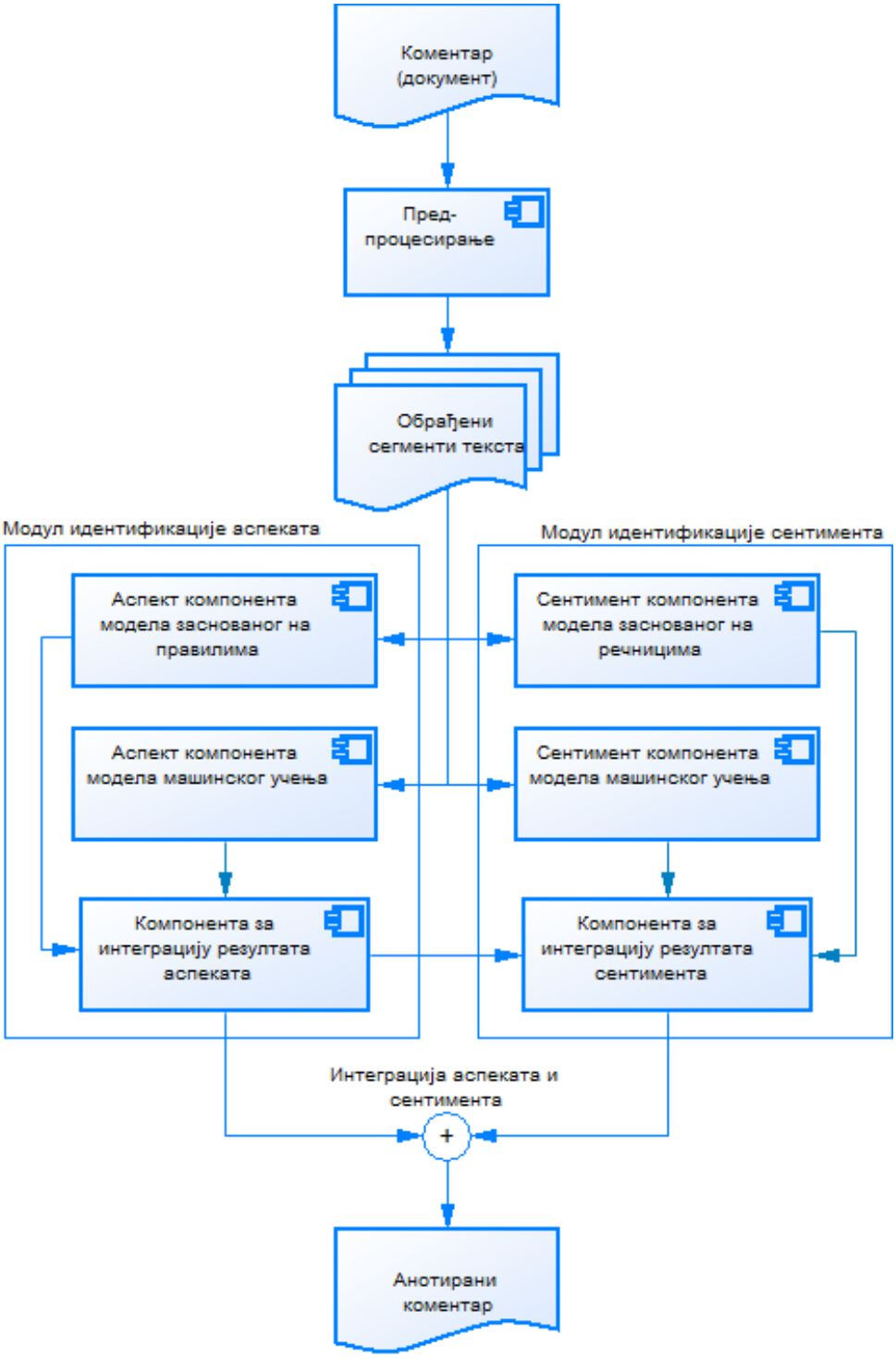
Систем за аутоматску анализу мишљења подразумева аутоматску анализу сегмената реченице (клаузе, фразе) при чему се одређује сентимент поларитет према одређеном аспекту студирања. Дакле, улаз у систем чине коментари студентских анкета а излаз из система анотације сентимент поларитета и аспекта за сваки сегмент сваке реченице коментара.

У наставку је описана методологија реализованог система за аутоматско издвајање мишљења студената из текстуалних коментара.

5.1 Методологија

Систем за анализу сентимента заснованом на аспектима се састоји од два корака – идентификацију аспеката студирања и доделу сентимент поларитета. Преглед система је дат на Слика 56.

Улаз у систем је коментар тј. документ који се прослеђује модулу за пред-процесирање података. Прво се врши дељење документа на реченице, затим реченице на сегменте текста. Сегменти текста у нашем случају представљају клаузе и фразе које се у овом систему третирају на исти начин. У наредном кораку се за сваки сегмент текста, у засебним модулима, врши анализа и идентификација аспеката и сентимент поларитета. Резултати се интегришу у крајњи резултат система који представља анотирани коментар са идентификованим аспектима и сентимент поларитетима.



Слика 56 *Преглед система*

5.2 *Пред-процесирање*

Обрада коментара подразумева уклањање знакова који у нашем случају немају семантички утицај на саму анализу текста. Из текста су уклоњени знакови попут специјалног карактера кодовања текста (енгл. *Byte Order Mark*), различитих врста заграда и знакова "[,],{,},(,), >, <, ~, _", интернет тагова и адреса "#,@,&,%" и осталих нестандардних *ASCII* знакова (од 128 до 255 знака). Емотикони имају значајну улогу у исказивању мишљења, поготову у коментарима писаних на интернету. Својим присуством допуњују мишљење и појачавају емоције исказаних у текстуалном коментару (Hogenboom *et al.*, 2013; Kralj Novak *et al.*, 2015; Wang and Castanon, 2015). Међутим, корпус К1 не садржи емотиконе с обзиром на то да су коментари писани руком (на папиру) док корпус К2 садржи мање од десет процената. Детаљнија анализа корпуса К2 указује на недоследну употребу емотикона у изразима мишљења (Kovačević *et al.*, 2020). Аутори истичу да су позитивни емотикони, поред допуне позитивног става, коришћени за ублажавање негативног става или појачавање ироније. Недоследност употребе емотикона указује на потребу додатног истраживања у контексту анализе сентимента и могућој сличности употребе у другим језицима. Због свега наведеног је одлучено да се у фази пред-процесирања врши уклањање емотикона из коментара. Након тога је извршена промена великих слова у мала слова и уклањање српских дијакритичких знакова чија употреба у оба корпуса није доследна.

Наредни корак обраде докумената укључује дељење коментара на реченице, па затим реченице на сегменте реченице (клаузе и фразе). За српски језик, према сазнању аутора дисертације, нема јавно доступних разделника реченица на клаузе и фразе, и стога је развијен наменски алат који је описан у наставку.

Алат за дељење коментара на сегменте реченице се састоји из два дела. У првом делу се врши раздвајање коментара на реченице засновано на анализи знакова интерпункције и великих слова. Изузетак су случајеви навођења редних и децималних бројева, скраћених префикса пре имена (нпр. проф. Ковачевић) и других скраћеница (нпр. ткз., тј., др., нпр., бр.). Међутим, ослањање само на знакове интерпункције приликом дељење реченица на сегменте не даје довољно добре резултате. Разлог је што се у српском језику често фразе и клаузе не налазе између знакова интерпункција у реченици него и између везника, које поред речи чине и изрази специфични за језик. На пример, „за разлику од", „у односу на", „с обзиром на" и тако даље. Из тог разлога се развијени алат за сегментацију реченице поред знакова интерпункције ослања и на присуство одређених врста речи попут везника, предлога и партиципа (глаголских прилога). С обзиром на недостатак јавно доступних алата за *POS* таговање за српски

језик током овог истраживања[22], присуство одређених врста речи детектовано је помоћу ручно прикупљених речника на основу обучавајућег скупа. Дељење реченица на сегменте се одвија према правилима описаним у наставку. Резултат примене правила на примеру реченица је дат у Табела XVI.

Табела XVI *Примери примене правила дељења реченица на сегменте*

Примери реченица и резултата примене правила	**Примена правила дељења**
Реченица: „Професор је фер и коректан **док** је асистент ужас благи**,** јако лоше ради задатке и на крају ништа не уради **а кад** треба да скида бодове ту је главни.“ Резултат: „Професор је фер и коректан“, „док је асистент ужас благи“, „јако лоше ради задатке и на крају ништа не уради“, „а кад треба да скида бодове ту је главни“.	Подела извршена на основу зареза, након тога на основу речи „док“ и израза „а кад“.
Реченица: „Асистенткиња је фина **и** професионална **иако** је организација на предмету чиста нула.“ Резултат: „Асистенткиња је фина и професионална“, „иако је организација на предмету чиста нула“	Подела извршена на основу речи „иако“ јер има предност у односу на реч „и“.
Реченица: „Асистент добро објашњава **и** добро се опходи према студентима**,** док је професор сушта супротност.“ Резултат: „Асистент добро објашњава“, „добро се опходи према студентима“, **„**док је професор сушта супротност“	Подела извршена на основу зареза, након тога на основу речи „и“. Дељење по речи „док“ је изостављено јер за резултат нису добијена бар два сегмента текста
Реченица: „Професор би требало да спреми питања која долазе на испит **и** да их студентима **с обзиром** да је литература преобимна **и** није нам потребно све.“ Резултат: „Професор би требало да спреми питања која долазе на испит и да их студентима“, „с обзиром да је литература преобимна и није нам потребно све“	Подела извршена на основу речи „с обзиром“ јер има предност у односу на речи „и“.

Правила се примењују у два корака. Прво се посматра постојање знака интерпункције тј. зареза, па се затим анализира присуство одређених врста речи и израза из прикупљених речника. Провера присуства речи и израза из речника се врши на основу приоритета. Већи приоритет је дат оној групи речи и израза који су анализом

[22] У време развоја алата није био активан РЕЛДИ *NLP* веб сервис за српски језик те стога није био укључен. Сада је доступан и биће укључен у будућем раду.

обучавајућег скуп корпуса К1 утврђени да се најчешће налазе између два сегмента реченице. На пример, речи и изрази попут “док”, “јер”, “за разлику од”, „у односу на“, „с обзиром“, „и мислим“, „иако“, „иначе“, „који је“, „пошто је“, „а кад“, имају већу предност од речи попут “али”, “или”, „а“, „и“, “па”. Након дељења на основу речи и израза из речника, дате речи и изрази остају део сегмента ради очувања контекста реченице.

Међутим, применом поменутих правила дељења реченице на сегменте је у одређеним случајевима дошло грешке. Након дељења су добијени мали сегменти са мањим бројем речи који су се показали да не носе информације о аспекту и сегменту (нпр. речи при набрајању одвојених зарезом и везником). Изузетак је случај да се у малом сегменту нађе баш реч аспекта која јасно указује на класу аспекта. Анализом добијених случајева се дошло до решења да се сегменти са мање од три речи придруже суседном сегменту. На овај начин су сачуване све информације тј. придружене речи дужим сегментима по броју речи ће употпунити информације о аспекту и сентимент поларитету. Прво се посматрају сегменти добијени у другом кораку дељења, а након тога се прелази на сегменте добијене у првом кораку. Сегменте са краја реченице се придружују претходном сегменту са три или више речи. Након тога се посматрају сегменти са почетка реченице и врши придруживање наредном сегменту са три или више речи. Примери специфичних случајева реченица су наведени у Табела XVII, а начин придруживања сегмената је описан у наставку.

Табела XVII *Специфични случајеви добијених сегмента реченице*

Примери реченица
Професор је као човек драг**,** насмејан **и** ведар.
Међутим**,** понекад предавања буду баш без везе.
Професор је изузетно коректан **и** професионалан**,** док је асистент страшно некоректан према студентима.

У првом примеру реченице Табела XVII, након примене оба корака дељења су добијени сегменти „Професор је као човек драг“, „насмејан“ и „ведар“. Последња два сегмента добијена у другом кораку дељења су сачињена од једне речи која не носе информације о аспекту студирања и сентимент поларитету, заједно. У овом случају је примењено решење спајања свих сегмената са краја реченице, који имају мање од три речи, претходном сегменту са више од три речи. Коначни добијени сегмент реченице након дељења је идентичан као и првобитна реченица. На исти начин, дељењем другог

примера реченице из Табела XVII се добијају два сегмента „Међутим“ и „понекад предавања буду баш без везе“, где ће се први сегмент придружити другом. У трећем примеру реченице Табела XVII, након првог корака дељења по зарезу су добијени сегменти „Професор је изузетно коректан и професионалан“ и „док је асистент страшно некоректан према студентима“. Даљим дељењем првог сегмента на основу речи „и“, други под сегмент „професионалан“ ће бити придружен првом под сегменту „Професор је изузетно коректан“. Дељење другог сегмента „док је асистент страшно некоректан према студентима“ на основу речи „док“ из речника неће бити извршено јер се за резултат не би добила два под сегмента. Поред предложеног решења спајања кратких сегмената по броју речи постоје случајеви у којима развијени алат не ради исправно. Анализа исправности развијеног алата је урађена на експериментелан начин.

Провера рада наменски развијеног алата за дељење коментара на сегменте реченица је извршена тако што су добијени сегменти реченица упоређивани са сегментима добијеним процесом анотације (Секција 4.1). Потпуним поклапањем добијених сегмената реченица и анотираних сегмената реченица је сматрано као исправан случај. Експеримент је извршен над свим анотираним сегментима из оба корпуса К1 и К2, а грешка је износила 0,68 процената за корпус К1 и 2,79 процената за корпус К2.

5.3 *Анализа аспеката*

Анализа аспеката је процес који сваком обрађеном сегменту текста додељује један аспект. Модул за анализу аспеката је сачињен од две компоненте, чији се резултати интегришу у један резултат према дефинисаној логици компоненте за интеграцију резултата. У наставку су описане обе компоненте модула за анализу аспеката – компоненте модела машинског учења и компоненте модела заснованог на правилима (Слика 56).

Током развијања система је експериментисано са више модела у оквиру модула а перформансе модела су приказане у Секцији 6.1.

5.3.1 Компонента модела машинског учења

Компонента машинског учења за анализу аспеката се састоји од три класификатора које чине алгоритми машинског учења. Први класификатор представља

вишекласни модел попут *NB*, *SVM* и *k-NN* (Секција 2.5.3). Други класификатор представља каскадни класификатор (енгл. *cascade classifier- CCLF*) који је сачињен из више вишекласних класификационих модела организованих у каскаду. Трећи класификатор чини модел дубоког учења који је један од најзаступљенијих у последњим истраживањима у *NLP* области а и шире. Појавом трансфера знања у моделима дубоког учења се проширила њихова примена и на случајеве са мањим скуповима података, попут овог истраживања. Иако су модели дубоког учења преузели примат у примени и постигнутим резултатима у *NLP* области у односу на стандардне моделе машинског учења, велика пажња у овом истраживању је посвећена ка стандардним моделима машинског учења из следећих разлога.

Према сазнању аутора ове дисертације, ово је један од првих радова везаних за аспектно базирану сентимент анализу на нивоу сегмента реченице за српски језик, па је иницијална фаза овог истраживања утврђивање граница перформанси стандардно коришћених модела машинског учења. Такође, аутори претходних истраживања у *ABSA* области (Mowlaei *et al.*, 2020; García-Díaz *et al.*, 2020; Al-Smadi *et al.,* 2018; Àlvarez-López *et al.,* 2016; Hercig *et al.,* 2016) су показали да постоје случајеви где стандардни модели машинског учења заједно са моделима заснованим на језичким ресурсима (правила, речници, онтологије, итд.) постижу боље перформансе у односу на поједине моделе дубоког учења. На основу горе наведеног и малог броја анотација за поједине аспекте (Табела XIV), прво је експериментисано са моделима заснованим на језичким ресурсима и стандардним моделима машинског учења. Након тога, је експериментисано са моделима дубоког учења са трансфером знања, који су се показали веома ефикасним у моделовању језика и разним *NLP* задацима.

Почетни корак примене стандардно коришћених модела машинског учења представља конструкција језичких особина. Конструкција језичких особина подразумева процес помоћу ког се текстуални сегменти текста трансформишу у вектор компоненти фиксне дужине. Након тога се дати вектори компоненти користе као део обучавајућег скупа алгоритама машинског учења. С обзиром на то да ово истраживање представља почетак експериментисања аспектно базиране сентимент анализе на нивоу сегмента текста за српски језик, коришћена је добро позната *BOW* техника репрезентације текста (Liu, 2015). Компоненте вектора при *BOW* репрезентацији текста су рачунате на основу фреквенције термина у односу на инверзну фреквенцију термина у документима, такозване *TF-IDF* статистике (Секција 2.2.1.3). У овом истраживању је експериментисано са више речника термина током обучавања модела, а њихова ефикасност, у погледу перформанси модела, тестирана на валидационом скупу података.

На основу резултата на валидационом скупу, утврђено је да уклањање стоп речи резултује погоршање перформанси модела машинског учења. Стога процес уклањања стоп речи није укључен при конструкцији језичких особина. Додатно су разматране три имплементације алата за стемовање за српски језик (Kešelj and Šipka, 2008; Milošević, 2012; РЕЛДИ) и једна за хрватски језик (Ljubešić *et al.*, 2007). Употреба алата за стемовање аутора (Milošević, 2012) је довела до побољшања перформанси модела машинског учења за седам процената па је из тог разлога овај алат за стемовање укључен у процес конструкције језичких особина. Анализом дужине *n-gram* термина је утврђено да су најбоље перформансе на валидационом скупу добијене узимањем у обзир *n-gram* термине дужине од један до четири речи. Такође, анализом је утврђено да уклањање 30 процената (9.960) најмање фреквентних *n-gram* термина од укупно 33.200 *n-gram* термина не утиче на перформансе модела. Коначан вектор језичких особина је имао укупно 23.240 *n-gram* термина.

Од стандардних модела машинског учења је експериментисано са класификаторима к најближих суседа (*k-NN*), машином потпорног вектора (*SVM*) и мултиномни наивни Бајес (*MNB*). За све наведене моделе је извршена оптимизација хипер-параметара применом насумичне методе (Pedregosa *et al.*, 2012), где су вредности параметара, у зависности од типа, биране на основу функције дистрибуције вредности или листе дискретних вредности. Перформансе модела током оптимизације хипер-параметара модела су рачунате на основу Ф-мере и петоструке унакрсне валидације на обучавајућем скупу. Модели су након тога евалуирани на валидационом скупу и најбоље резултате је остварио *SVM* модел са линеарним језгром, који је онда коришћен и у каскадном класификатору.

Каскадни класификатор чини скуп *SVM* класификатора организованих у каскадну структуру по узору на хијерархију анотационе шеме аспеката (Секција 4.1). Сваки *SVM* класификатор је обучен за вршење бинарне класификације. На пример: класификатор *Наставник-Предмет* врши класификацију на класе аспекта *Наставник* и *Предмет*. Процес класификације почиње од врха хијерархије и наставља се докле год се не додели један од аспеката најниже хијерархије. Предност оваквог приступа је модуларност, где сваки од класификатора може бити независно обучен и оптимизован за дати задатак класификације. Такође, сви бинарни класификатори каскадне структуре не морају бити истог типа класификатора машинског учења.

Од вишејезичних пре-обучених модела дубоког учења је експериментисано са *BERT*, *XLM* и *XLM-RoBERTa* моделима који подржавају српски језик (Секција 2.2.1.3.1).

Коришћена је библиотека компаније *HuggingFace*[23] која је под исти интерфејс интегрисала преко 1.700 претходно обучених модела за разне *NLP* задатке попут класификације текста, машинског превођења текста, сумирања текста, аутоматског одговарања на питања, и тако даље. Поред обучених модела библиотека садржи и алгоритме за токенизацију текста, оптимизацију и планирање обучавања модела.

С обзиром на то да су *BERT*, *XLM* и *XLM-RoBERTa* модели обучени на различитим корпусима на више језика, коришћени су токенизатори који одговарају датим моделима. Применом токенизатора су трансформисане речи улазних секвенци у низове идентификатора речи у речнику токенизатора. Затим су модели дообучавани на задатку вишекласне класификације применом обучавајућег скупа. При чему је дообучаван само последњи слој модела који је задужен за класификацију. Вредности хипер-параметара модела су оптимизоване применом Бајесове оптимизације (Секција 2.5.3.2.2.3.1).

Техничка спецификација коришћених модела и оптимизоване вредности хипер-параметара коришћених при дообучавању су наведени у наставку. Коришћен је основни вишејезични *BERT* модел који је претходно обучен на корпусу од 104 језика, док су *XLM* и *XLM-RoBERTa* модели обучени на корпусу од 100 језика. Речници токенизатора *BERT*, *XLM* и *XLM-RoBERTa* модела садрже преко 110.000, 200.000 и 250.000 речи, респективно. Обучавајући скуп података је подељен на подскупове од два примера и модел је дообучаван на четири епохе. Коришћен је корак учења од $3e^{-5}$ и *Adam* алгоритам оптимизације параметара модела (Секција 2.5.3.2.2.3). Током обучавање је коришћена паралелизација обраде података на процесорима графичке картице. Тиме је процес обучавања модела убрзан у просеку двадесет седам пута по епохи (са четири сата на девет минута по епохи). Коришћено је *Google Colab* окружење[24] на дељеном серверу са графичком картицом *NVIDIA Tesla T4* са 2.560 језгара и 16 гигабајта *GDDR6* меморије.

5.3.2 Компонента модела заснованог на правилима

Поред компоненте машинског учења коришћен је модел заснован на правилима, представљен унутар своје засебне компоненте. Анализом резултата компоненте машинског учења на валидационом скупу, уочено је да постоји простор за побољшање прецизности и одзива модела. С обзиром на специфичну терминологију коришћену у корпусу К1, постоје мање фреквентни термини који јасно указују на одређени аспект.

[23] Компанија *HuggingFace*: https://huggingface.co/
[24] *Google Colab*: https://colab.research.google.com/

Правила су формирана ручно у облику регуларних израза и обухватају изразе који се најчеше јављају у домену високог образовања. Ослањају се на ручно прикупљене речнике израза и врста речи (придева, прилога, глагола, именица, речца, везника). Укупно је сачињено четрдесет правила за седам аспеката студирања (Прилог А). У Табела XVIII је дат број сачињених правила по аспектима студирања.

Табела XVIII *Број сачињених правила по аспектима*

Аспект	Број правила
Наставник	8
Настава	6
Однос	12
Предмет	2
Материјали	3
Организација	7
Остало	2

Правила се могу међусобно преклапати по аспектима тј. активирати правила различитх аспеката за исти сегмент текста. У том случају се преузима специфичнији аспект по узору на хијерархијску шему анотације (Слика 54). Тачније, правила су примењена по следећем редоследу приоритета – *Наставник*, *Настава*, *Однос*, *Предмет*, *Материјали*, *Организација*, *Остало*. На пример, уколико се за исту секвенцу активира правило за аспект *Наставник* и *Настава*, у том случају се додељује аспкет *Настава*. Употреба правила довела је до повећања Ф-мере за три процента на валидационом скупу. Примери правила су дати у Табела XIX.

Табела XIX *Примери правила заснованих на регуларним изразима.*

Напомена: "..." у правилима замењују више термина речника

Примери	**Аспекти**	**Правила**
наставник је изузетан човек	Наставник	(\s|^)((nastavn\w*)[ ](je|su)[ ](fenomenal|supe|izuzet|dobar|losx|\w*spor|interesant|najbolj|izuzet|pozitiv|negativ|...)\w*)?[ ]((cxovek|legend|osob)\w*)?
наставник је спреман да помогне	Однос	(\s|^)((nastavn\w*)[ ](je|su)[ ](sprem|rad|volj|raspolozx|...)\w*)(\w+)?[ ]((pomogn|saslusz|(izadj\w*)(u susret)|saradnj)\w*)
али је наставник лош предавач	Настава	(\s|^)((nastavn\w*)[ ](je|su)[ ](fenomenal|supe|izuzet|dobar|losx|\w*spor|interesant|najbolj|izuzet|pozitiv|negativ|...)\w*)?[ ]((predavacx| strucxnja|profesional)\w*)

5.3.3 Интеграција резултата

Резултати анотација аспеката модела обе компоненте се интегришу у компоненти за интеграцију резултата аспеката (Слика 56). Применом модела машинског учења и модела заснованог на правилима, сваки сегмент добија најмање три а највише четири анотације (у случају активирања правила) за категорију аспект. С обзиром на то да постоји могућност да се додељене анотације међусобно разликују, експериментисано је са више стратегија интеграција модела за разрешење потенцијалних конфликата. Стратегије су формиране на основу експеримената на валидационом скупу и тичу се рангирања модела према одређеном редоследу. Одабрана је стратегија интеграције модела где се појединачно интегришу две групе модела, група класичних модела и група модела дубоког учења.

У групи класичних модела се интегришу анотације модела заснованог на правилима, каскадног модела и класичног модела машинског учења. Потенцијални конфликти се решавају тако што се даје предност моделу заснованом на правилима а затим каскадном моделу па класичном моделу машинског учења. Резултат миграције су сегменти текста којима су додељене тачно по једна класа аспекта. У групи модела дубоког учења се интегришу анотације модела дубоког учења где се потенцијални конфликти решавају на основу приоритета по класама. Приоритети су дефинисани на основу перформанси модела на валидационом скупу за одређену класу аспекта.

Резултат интеграције ове групе модела су такође сегменти текста којима су додељене тачно по једна класа аспекта. Уколико је потребно, интеграција модела две групе у један крајњи модел се врши на основу приоритета по класама.

Одабрана стратегија интеграције модела не гарантује најбоље укупне перформансе на скупу за тестирање у односу на појединачне интеграционе моделе. Разлог тога је то што је стратегија интеграције модела формирана на скупу за валидацију а не скупу за тестирање.

5.4 Анализа поларитета сентимента

Сваком сегменту текста коме је додељен један аспект, додељује један сентимент. Развијени су модели само за позитиван и негативан сентимент јер је фреквенција класа неутралног сентимента веома мала (до два процента у оба корпуса). Сентимент модел је, попут аспектног модела, сачињен од две компоненте чији резултати се интегришу према задатој логици. У наставку су описане обе компоненте модула за анализу сентимента – компоненте модела машинског учења и компоненте модела заснованог на речницима (Слика 56).

Током развијања система је експериментисано са више модела у оквиру модула а перформансе модела су приказане у наредној Секцији 6.2.

5.4.1 Компонента модела машинског учења

Компонента машинског учења сентимент модула се састоји од два модела класификатора. Први модел класификатора је по начину формирања идентичан моделу машинског учења из модула за идентификацију аспеката а обучен је за две сентимент класе – позитиван и негативан. Од модела машинског учења је експериментисано са истим моделима као код класификатора машинског учења у модулу за идентификацију аспеката. За све наведене моделе је извршена оптимизација хипер-параметара, на начин описаним у аспект модулу, а перформансе сентимент анализе су измерене над валидационим скупом. Други модел класификатора је такође по начину формирања идентичан моделу дубоког учења у модулу за идентификацију аспеката. Коришћени су вишејезични *BERT, XLM* и *XLM-RoBERTa* модели дубоког учења који су претходно обучени на корпусима на више језика. Дообучавани су на исти начин као у модулу за идентификацију аспеката за задатак бинарне класификације.

5.4.2 Компонента модела заснованог на речницима

Друга компонента представља модел заснован на речницима позитивних и негативних речи, а и речника инкремента и декремента, описаних у наставку. За сваки сегмент текста врши се аутоматска детекција присуства једног или више термина из сва четири речника. На основу препознатих термина рачуна се укупан сентимент сегмента текста по поступку аутора (Liu, 2015) описаном у Секцији 3.1. Сваки детектовани инкремент и позитван израз повећавају укупан резултат за један, док декремент и негативан израз смањују за један. На пример, сегмент "Наставник је максимално посвећен студентима" садржи позитивну реч „посвећен" и реч „максимално" која појачава интезитет позитивне речи. Дакле укупан резултат овог сегмента је два, што значи да је аутоматски додељен позитиван сентимент. Описани приступ је сличан приступима који су описани у радовима *Hu and Liu* (2004) и *Kim and Hovy* (2007).

У оквиру истраживања аутора (Grljević, 2016) су током анотирања корпуса К2, поред аспекта и сентимент поларитета, анотиране и сентимент речи или изрази које носе позитиван или негативан сентимент. Аутори су анотиране сентимент речи и изразе груписали у сентимент речнике по поларитету. Додатно су вођене белешке о свим језичким специфичностима које могу бити занимљива у погледу анотираног сентимента. На датих белешки су формирани речници интезификатора, неутрализатора и негације сентимента.

Речници интезификатора садрже термине који повећавају интензитет сентимента попут придева и прилога. На пример, речи „најбољи", „одличан" и „диван" имају позитиван сентимент поларитет. Уколико се у сегменту текста налази позитиван сентимент и још један термин интезификатора, онда ће при сентимент анализи позитиван сентимент имати појачан интензитет. Насупрот речнику интезификатора је речник неутрализатора који садржи термине који на сличан начин смањују интензитет сентимента. На пример, речи попут „безобразан", „разочаран" и „намргођен" имају негативан сентимент поларитет. Речник негације садржи речи и термине који мењају сентимент поларитет у једном сегменту текста. На пример, уколико се испред позитивне речи „свиђати" дода реч негације „не", крајњи поларитет је негативног сентимента. У оквиру модела заснованог на речницима је експериментисано са свих пет наведених речника.

5.4.3 Интеграција резултата

Резултати анотација сентимент модела обе компоненте се интегришу у компоненти за интеграцију резултата сентимента (Слика 56) која ради на сличном принципу као и код аспектне анализе (Секција 5.3.3). Применом модела машинског учења и модела заснованог на речницима, сваки сегмент добија најмање два а највише три анотације (у случају активирања речника) за категорију сентимент. Потенцијални конфликти модела се решавају према стратегији интеграције модела која је описана у наставку.

Стратегија интеграције модела се примењује на исте две групе модела, група класичних модела и група модела дубоког учења. У групи класичних модела се интегришу анотације модела заснованог на речницима и класичног модела машинског учења. Потенцијални конфликти се решавају према приоритету модела који зависи од резултата интеграције модела аспеката. Анализом је откривено да модели засновани на речницима дају значајно боља предвиђања за аспекте првог нивоа каскаде (*Наставник* и *Предмет*) у поређењу са аспектима другог нивоа (Слика 54). Стога, у случају детекције аспеката првог нивоа каскаде, приоритет се даје анотацијама модела заснованог на речницима. За све остале аспекте се даје предност класичном моделу машинског учења. Резултат миграције су сегменти текста којима су додељене тачно по једна класа сентимента. У групи модела дубоког учења се интегришу анотације модела дубоког учења где се потенцијални конфликти решавају на основу приоритета по класи сентимента. Већи приоритет има модел који је остварио боље перформансе на валидационом скупу. Резултат миграције групе модела дубоког учења су такође сегменти текста којима су додељене тачно по једна класа сентимента. Уколико један од интегрисаних модела две групе није постигао најбоље перформансе по свим класама сентимента онда се приступа интеграцији модела две групе у један крајњи модел. У том случају се интеграција врши на основу приоритета по класама.

Одабрана стратегија интеграције модела не гарантује најбоље укупне перформансе на скупу за тестирање у односу на појединачне интеграционе моделе. Разлог тога је то што је стратегија интеграције модела формирана на скупу за валидацију а не скупу за тестирање.

6 Експериментални резултати и дискусија

У циљу евалуације перформанси предложене методологије аспектно базиране сентимент анализе, експерименти су извршени на корпусу К1 златног стандарда. Корпус К1 је подељен на подскупове за обучавање, валидацију и тестирање (Секција 2.4). Ради додатне провере квалитета система, извршена је евалуација и на корпусу К2 (Секција 4). Експерименти су вршени на свим моделима описаним у методологији, а перформансе модела приказане по *ABSA* задацима - задатак идентификације аспеката и задатак идентификације сентимент поларитета. Поред перформанси појединачних модела су приказане перформансе најбољих интеграционих модела по *ABSA* задацима добијених на основу стратегија интеграције модела (Секције 5.3.3, 5.4.3). Перформансе модела су мерене прецизношћу, одзивом и тежинској макро Ф-мери (Секција 2.4.1).

У наставку су дате перформансе модела задатка идентификације аспеката и задатка идентификације сентимент поларитета. Након тога је дата анализа грешака најбољих интеграционих модела. На крају секције је дата дискусија импликације резултата и ограничења овог истраживања.

6.1 Задатак идентификације аспеката

Резултати задатка идентификације аспеката појединачних модела и најбољег интеграционог модела групе класичних модела су представљени у Табела XX. Разматрајући моделе засебно, *SVM* модел је постигао најбоље перформансе за оба корпуса (Ф-мера од 0,76 и 0,67), док примена каскадног модела (*CCLF*) је резултовала у побољшању Ф-мере од један проценат. Додатак компоненте базиране на правилима је повећала Ф-меру *SVM* модела за два процента и каскадног модела за три процента за корпус К1, и за један проценат за корпус К2. Свеукупно најбоље перформансе су постигнуте интеграцијом компоненте засноване на правилима и *SVM* модела (*AM1*) и каскадног модела (*AM2*) у модел *AM3* према стратегији описаној у Секцији 5.3.3.

Табела XX *Резултати идентификације аспеката појединачних модела и најбољег модела интеграције групе класичких модела*

МОДЕЛ	КОРПУС					
	К1			К2		
	P	***R***	***F***	***P***	***R***	***F***
k-NN	0,65	0,65	0,64	0,63	0,63	0,60
NB	0,73	0,72	0,70	0,65	0,65	0,60
SVM	0,78	0,76	0,76	0,67	0,70	0,67
CCLF	0,78	0,77	0,77	0,68	0,69	0,68
AM1 (SVM + RB)	0,80	0,79	0,78	0,69	0,70	0,68
AM2 (CCLF + RB)	0,81	0,80	0,80	0,70	0,69	0,69
AM3 (AM1 + AM2)	0,81	0,81	**0,81**	0,71	0,71	**0,71**

Напомена: P – прецизност, R – одзив, F – Ф-мера

Резултати идентификације аспеката приказаних по класама аспеката за модел *AM3* су дати у Табела XXI. Као што се може и видети у табели, Ф-мере већине аспеката за корпус К1 су у опсегу од 0,80 ди 0,89. За аспекте *Организација*, *Предмет* и *Настава* су остварене мање Ф-мере редом од 0,81, 0,80 и 0,77.

Табела XXI *Резултати идентификације аспеката приказаних по аспектима најбољег модела интеграције групе класичких модела*

АСПЕКТ / МОДЕЛ	КОРПУС							
	К1				К2			
	P	***R***	***F***	***FREQ***	***P***	***R***	***F***	***FREQ***
Наставник	0,91	0,86	0,89	121	0,77	0,87	0,83	976
Настава	0,71	0,82	0,77	188	0,63	0,55	0,58	302
Однос	0,79	0,92	0,85	24	0,70	0,60	0,64	482
Предмет	0,78	0,82	0,80	39	0,50	0,46	0,49	24
Материјали	0,84	0,92	0,88	78	0,71	0,49	0,58	61
Организација	0,82	0,80	0,81	269	0,67	0,06	0,12	31
Остало	1,00	0,31	0,49	32	0,00	0,00	0,00	29
AM3	0,81	0,81	**0,81**	751	0,71	0,71	**0,71**	1.905

Напомена: P – прецизност, R – одзив, F – Ф-мера, FREQ - фреквенција

Резултати задатка идентификације аспеката појединачних модела и најбољег интеграционог модела групе модела дубоког учења су представљени у Табела XXII. Разматрајући перформансе засебних модела дубоког учења, *XLM* и *XLM-RoBERTa* модели су постигли боље перформансе над класама аспеката у односу на *BERT* модел за оба корпуса на тест скупу. Перформансе *XLM* и *XLM*-RoBERTa модела се разликују по класама аспеката, стога су најбоље перформансе за оба корпуса (Ф-мера од 0,76 и 0,76) постигнуте моделом *AM4* који чини интеграцију *XLM-RoBERTa* и *XLM* модела.

Интеграција модела је заснована према описаној стратегији интеграције модела на валидационим скупом (Секција 5.3.3). Тако су за корпус К1 анотације за класе *Наставник* и *Предмет* добијене *XLM-RoBERTa* моделом, док су анотације за преостале класе аспеката (*Настава*, *Однос*, *Материјали*, *Организација*, *Остало*) добијене *XLM* моделом. Док су за корпус К2 анотације за класе *Наставник*, *Настава* и *Однос* добијене *XLM-RoBERTa* моделом, док су анотације за преостале класе аспеката (*Предмет*, *Материјали*, *Организација*, *Остало*) добијене *XLM* модела. Поређењем перформанси најбољег модела из првог корака стратегије интеграције модела (Табела XX) се може уочити да су Ф-мере *SVM* модела и *AM4* модела сличне.

Табела XXII *Резултати идентификације аспеката појединачних модела и најбољег модела интеграције групе модела дубоког учења*

МОДЕЛ	КОРПУС					
	К1			К2		
	P	***R***	***F***	***P***	***R***	***F***
BERT	0,76	0,74	0,75	0,67	0,71	0,69
XLM	0,75	0,74	0,75	0,72	0,73	0,72
XLM-RoBERTa	0,75	0,74	0,74	0,74	0,76	0,75
AM4 (*XLM+XLM-RoBERTa*)[25]	0,76	0,75	**0,76**	0,76	0,77	**0,76**
Напомена: P – прецизност, R – одзив, F – Ф-мера						

Резултати идентификације аспеката приказаних по категорији аспеката за модел *AM4* су дати у Табела XXIII. Као што се може и видети у тебели, Ф-мере већине аспеката за корпус К1 су у опсегу од 0,70 до 0,83. Међу аспектима са мањом Ф-мером је аспект *Предмет* (Ф-мера од 0,70), који уз аспекте *Настава* и *Организација* (Ф-мере редом од 0,71 и 0,76).

У случају корпуса К2, резултати најбољег модела интеграције групе модела дубоког учења су свеукупно изједначени са моделом корпуса К1 али и значајно виши у односу на најбољи модел интеграције групе класичких модела (*AM3*). На основу резултата се може уочити знатно побољшање перформанси за аспекте Настава, Однос и Материјали, у односу на модел *AM3*.

[25] Напомена: Интеграција модела је заснована на валидационом скупу где су перформансе *XLM* и *XLM-RoBERTa* модела биле боље за све класе аспеката у односу на *BERT* модел

Табела XXIII *Резултати идентификације аспеката приказаних по аспектима најбољег модела интеграције групе модела дубоког учења*

АСПЕКТ / МОДЕЛ	КОРПУС							
	К1				К2			
	P	***R***	***F***	***FREQ***	***P***	***R***	***F***	***FREQ***
Наставник	0,83	0,83	0,83	121	0,82	0,88	0,85	976
Настава	0,65	0,76	0,71	188	0,71	0,73	0,72	302
Однос	0,75	0,84	0,79	24	0,77	0,69	0,72	482
Предмет	0,69	0,72	0,70	39	0,35	0,46	0,40	24
Материјали	0,83	0,83	0,83	78	0,67	0,61	0,64	61
Организација	0,83	0,68	0,76	269	0,22	0,13	0,16	31
Остало	0,49	0,61	0,54	32	0,00	0,00	0,00	29
AM4	0,76	0,75	**0,76**	751	0,76	0,77	**0,76**	1.905

Напомена: P – прецизност, R – одзив, F – Ф-мера, FREQ - фреквенција

За корпус К1, *AM3* модел није постигао боље перформансе од *AM4* модела за сваку класу аспекта виског образовања на валидационом скупу. Док за корпус К2, *AM4* модел није остварио најбоље резултате на валидационом скупу код свих класа аспеката у поређењу са најбољим моделом интеграције групе класичких модела (*AM3*).

У том случају је крајњи модел за идентификацију аспеката *AM5* добијен интеграцијом *AM3* и *AM4* модела (Секција 5.3.3). За корпус К1, анотације *AM5* модела за аспект *Материјали* су добијене *AM4* моделом а анотације за остале аспекте *AM3* моделом. У случају корпуса К2, анотације *AM5* модела за аспекте *Предмет* и *Остало* су добијене моделом *AM3* а анотације за остале аспекте *AM4* моделом. Перформансе крајњег *AM5* модела за оба корпуса су приказане у Табела XXIV.

Табела XXIV *Резултати идентификације аспеката приказаних по аспектима - крајњи модел*

АСПЕКТ / МОДЕЛ	КОРПУС							
	К1				К2			
	P	***R***	***F***	***FREQ***	***P***	***R***	***F***	***FREQ***
Наставник	0,91	0,86	0,89	121	0,82	0,88	0,85	976
Настава	0,71	0,82	0,77	188	0,71	0,73	0,72	302
Однос	0,79	0,92	0,85	24	0,77	0,69	0,72	482
Предмет	0,78	0,82	0,80	39	0,50	0,46	0,49	24
Материјали	0,83	0,83	0,83	78	0,67	0,61	0,64	61
Организација	0,82	0,80	0,81	269	0,22	0,13	0,16	31
Остало	1,00	0,31	0,49	32	0,00	0,00	0,00	29
AM5 (AM3 + AM4)	0,81	0,80	**0,80**	751	0,76	0,77	**0,76**	1.905

Напомена: P – прецизност, R – одзив, F – Ф-мера, FREQ - фреквенција

Међутим, укупне перформансе *AM5* модела нису надмашиле перформансе модела *AM3* модела на тест скупу података корпуса К1. Просечна Ф-мера *AM5* модела је за један проценат (0,01) лошија у односу на просечну Ф-меру *AM3* модела. У случају корпуса К2, *AM5* модел је постигао боље перформансаме у односу на модел *AM3*. Тачније, просечна Ф-мера *AM5* модела од 0,76, колико износи и просечна Ф-мера *AM4* модела, је за пет процената боља од просечне Ф-мере *AM3* модела.

6.2 *Задатак идентификације сентимента*

Резултати задатка идентификације сентимента појединачних модела и најбољег интеграционог модела групе класичних модела су представљени у Табела XXV. Од појединачних модела машинског учења, *SVM* модел је постигао најбоље резултате са Ф-мером од 0,89 за корпус К1 и 0,78 за корпус К2. Као што је и очекивано, модел заснован на речницима је имао знатно боље резултате на корпусу К2, на основу ког су речници и настали. Интеграција *SVM* модела са моделом заснованом на речницима, где су конфликти решавани давањем приоритета моделу заснованом на речницима, резултирао је смањењем Ф-мере. Овај модел резултовао је са највишом постигнутом Ф-мером од 0,90 за корпус К1 и 0,81 за корпус К2.

Табела XXV *Резултати идентификације сентимента појединачних модела и најбољег модела интеграције групе класичких модела*

МОДЕЛ	**КОРПУС**					
	К1			**К2**		
	P	***R***	***F***	***P***	***R***	***F***
k-NN	0,86	0,89	0,87	0,73	0,74	0,73
NB	0,88	0,90	0,88	0,77	0,78	0,77
SVM	0,88	0,90	0,89	0,78	0,79	0,78
DB	0,80	0,55	0,57	0,79	0,79	0,79
SM1 (*SVM* + *DB*)	0,81	0,66	0,69	0,79	0,76	0,77
SM2 (*SVM* + *DB**)	0,89	0,90	**0,90**	0,81	0,81	**0,81**

Напомена: P – прецизност, R – одзив, F – Ф-мера

Резултати идентификације сентимента приказаних по сентимент поларитету за модел *SM2* су дати у Табела XXVI. Као што је и очекивано, Ф-мера је значајно већа за поларитете где је сентимент више фреквентан у оба корпуса. Анотације негативног поларитета чине више од 70 процента за корпус К1, док анотације позитивног поларитета чине више од 60 процента корпуса К2.

Табела XXVI *Резултати идентификације сентимента приказаних по сентименту најбољег модела интеграције групе класичких модела*

СЕНТИМЕНТ / МОДЕЛ	КОРПУС							
	К1				К2			
	P	R	F	FREQ	P	R	F	FREQ
Негативан	0,91	0,97	0,94	585	0,73	0,81	0,77	773
Позитиван	0,86	0,78	0,83	165	0,86	0,82	0,84	1.116
SM2	0,89	0,90	**0,90**	750	0,81	0,81	**0,81**	1.889

Напомена: P – прецизност, R – одзив, F – Ф-мера, FREQ - фреквенција

Резултати задатка идентификације сентимента појединачних модела и најбољег интеграционог модела групе модела дубоког учења су представљени у Табела XXVII. Разматрајући засебно моделе дубоког учења, *XLM-RoBERTa* модел је постигао најбоље перформансе за оба корпуса.

Табела XXVII *Резултати идентификације сентимента појединачних модела и најбољег модела интеграције групе модела дубоког учења*

МОДЕЛ	КОРПУС					
	К1			К2		
	P	***R***	***F***	***P***	***R***	***F***
BERT	0,94	0,94	0,94	0,80	0,80	0,80
XLM	0,93	0,93	0,93	0,84	0,84	0,84
XLM-RoBERTa	0,96	0,96	**0,96**	0,90	0,90	**0,90**

Напомена: P – прецизност, R – одзив, F – Ф-мера

Резултати сентимент идентификације приказаних по поларитету сентимента за модел *XLM-RoBERTa* су приказани у Табела XXVIII. Из истог разлога као и код модела *SM2*, Ф-мере су значајно веће за поларитете где је сентимент више фреквентан у оба корпуса. Као што се може и видети у тебели, Ф-мера за негативан сентимент поларитет за корпус К1 је 0,97 док је за позитиван поларитет за корпус К2 Ф-мера износи 0,91. Са укупном Ф-мером од 0,96 за корпус К1 и 0,90 за корпус К2, модел *XLM-RoBERTa* је постигао боље перформансе у односу на *SM2* модел.

Табела XXVIII *Резултати идентификације сентимента приказаних по сентименту најбољег модела интеграције групе модела дубоког учења*

СЕНТИМЕНТ / МОДЕЛ	КОРПУС							
	К1				К2			
	P	R	F	FREQ	P	R	F	FREQ
Негативан	0,97	0,97	0,97	585	0,87	0,89	0,88	773
Позитиван	0,91	0,91	0,91	165	0,92	0,90	0,91	1.116
XLM-RoBERTa	0,96	0,96	**0,96**	750	0,90	0,90	**0,90**	1.889

Напомена: P – прецизност, R – одзив, F – Ф-мера, FREQ - фреквенција

С обзиром на то да је интеграциони модел групе модела дубоког учења остварио најбоље перформансе, онда није било потребе за интеграцијом модела две групе у један крајњи модел (Секција 5.4.3). Коначни модел идентификације сентимент поларитета је *XLM-RoBERTa*.

6.3 Анализа грешака

У наставку је описана анализа грешака интеграционих модела за идентификацију аспеката и сентимент поларитета. Крајњи модел за идентификацију аспеката чине анотације интеграционих модела две групе модела – класичних модела и модела дубоког учења. Док крајњи модел за идентификацију сентимент поларитета чини интеграциони модел групе модела дубоког учења.

Анализом грешака интеграционог модела групе класичних модела за задатак идентификације аспеката (Табела XXI) је утврђено да је већина погрешно негативних (Секција 2.4.1) анотација за аспект *Настава* погрешно класификована као аспект *Организација*, док је већина погрешно позитивних (Секција 2.4.1) анотација било означено са апсектом *Организација* у златном стандарду. Такође, за класу *Предмет* је утврђено да више од пола погрешно негативних анотација чине погрешне класификације као аспект *Организација* а мање од 30 процената погрешно негативних анотација погрешне класификације као аспект *Настава*.

Даља истраживања су показала да је конфузија између наведена три аспекта био резултат преклапања високо рангираних *TF-IDF* термина између поменутих аспеката. Наиме, у коментарима анотираним класом *Организација* се коментарише начин организовања предмета и предметних обавеза и предлаже боља организација истих. На сличан начин се коментарише извођење наставе од стране наставника, те се исти термини појављују и у коментарима анотираних класом *Настава*. На пример,

реченицом „Требало би да имамо два уместо једног колоквијума“ се коментарише организација предмета, док се реченицом „Требао је данас да вежбамо за колоквијум“ коментарише настава. Из истог разлога се обе класе *Организација* и *Настава* налазе у већини погрешно негативних анотација класе *Предмет*, која узгред има малу фреквенцију у корпусу К1.

Аспект са најмањом Ф-мером је био аспект *Остало* због велике варијабилности његових анотација и ниске фреквенције појављивања у корпусима (Табела XIV). Аспект *Остало* обухвата велики број лексички варијабилних анотација које се не односе на домен високог образовања и које такође имају малу фреквенцију понављања у корпусу. На пример, „Све најбоље“ „Све је ОК“, „Спава ми се“, „Супер сте“. Због ових разлога, модели машинског учења који се ослањају на *BOW* репрезентацију нису били у стању да успешно идентификују аспект *Остало*.

У случају корпуса К2, резултати најбољег модела интеграције групе класичких модела су свеукупно лошији него у случају К1. Иако је благи пад у перформансама био очекујући у поређењу са К1 због компоненте засноване на правилима, која је развијена и прилагођена корпусу К1. Даљом анализом је откривено да се укупни пад перформанси на корпусу К2 може објаснити начином на који се формирају подаци за обучавање *ML* модела. Сваки пример у обучавајућем скупу је био један сегмент реченице (фраза или клауза) који је добијен на основу аутоматског разделника описаног у Секцији 5.2. Лабеле сегмента текста након дељења су одређене на основу поклапања са анотацијама златног стандарда. На пример, анотирана реченица из златног стандарда изгледа овако:

„*<Nastavnik> Asistent je veoma profesionalan </Nastavnik>, <Organizacija> međutim predmet je lose organizovan </Organizacija>*“

Анотирани сегменти текста су подвучени у примеру изнад. Након примене разделника су добијена два сегмента: „*Asistent je veoma profesionalan*“ и „*međutim predmet je lose organizovan*“. Затим је у обучавајући скуп додат први сегмент са додељеним аспектом *Наставник* и други сегмент са додељеним аспектом *Организација*. Како анотатори нису били ограничени експлицитним анотирањем сегмената реченица (фраза и клауза), анотације нису морале да се подударају са границама фраза и клауза. Неслагања између анотираног сегмента текста је било далеко веће за корпус К2 него за корпус К1. У случају корпуса К2, 4,28 процената анотација се није подударало са фразама (или клаузама), док је ово био случај у само 0,68 процента за корпус К1. На пример, реченица златног стандарда:

“*<Odnos> Veoma korektan odnos prema studentima, ohrabruje ih da razmišljaju svojom glavom, a ne da samo da uče činjenice napamet iz literature. </Odnos>*”

Реченица је подељена у три сегмента (сегменти су подвучени у примеру). Сва три сегмента су укључена у обучавајући скуп као три одвојена примера са аспектом *Однос*. Док прва два сегмента садрже термине (појмове) који су интуитивно повезани са аспектом *Однос* („*korektan odnos*“, „*ohrabruje ih*“), трећи сегмент садржи термине „*činjenice*“ и „*literature*“ (који су више типични за аспект *Материјали*) што ће, због *BOW* репрезентације, бити погрешно повезане са аспектом *Однос*.

Анализом грешака интеграционог модела групе модела дубоког учења за задатак идентификације аспеката (Табела XXIII) је утврђено да 60 процената погрешно негативних анотација за аспект *Настава* су погрешно класификовани као аспект *Организација*, док је 58 процената погрешно позитивних анотација чинило апсект *Организација*. Такође, утврђено је да у веома сличној размери погрешних анотација за аспект *Организација* чини аспект *Настава*. У 58 процената погрешно негативних анотација за аспект *Организација* су погрешно класификовани као аспект *Настава*, док је 53 процената погрешно позитивних анотација чинило апсект *Настава*. За аспект *Предмет* је утврђено да 70 процената погрешно негативних анотација чине погрешне класификације аспекта *Организација*, док је 43 процената и 29 процената погрешно позитивних анотација чинило аспект *Организација*, односно аспект *Настава*. Аспект са најмањом Ф-мером је био аспект *Остало* због велике варијабилности његових анотација и ниске фреквенције појављивања у корпусима (Табела XIV). Добијена Ф-мера за аспект *Остало* је боља у односу на модел *AM3* из разлога постојања великог речника језичког модела који је претходно обучен на великим корпусима (нпр. 2,5 терабајта података). У случају класификације текста, када се појави реч која није била присутна у обучавајућем скупу, класични модели машинског учења попут *SVM* ће дату реч занемарити и тиме изгубити информацију. Док језички модели попут *BERT*-а користе велике речнике речи и под речи што смањује вероватноћу губитка информација.

Анализом грешака интеграционог модела групе класичних модела за задатак идентификације сентимента (Табела XXVI) је утврђено да је преко 60 процената погрешно позитивних анотација било означено позитивним сентиментом због појављивања речи ван речника модела машинског учења. На пример, сегмент текста „Асистент неуверљив и спетљан“ је у *BOW* приступу посматран само као реч „асистент“ јер се придеви „неуверљив“ и „спетљан“ нису појављивали током обучавања модела. А реч „асистент“ је високо рангиран *TF-IDF* термин у позитивним коментарима. Такође, преко 30 процената погрешно позитивних анотација се може приписати губљењу контекста у *BOW* приступу. На пример, сегмент текста „Једино бих волео да су вежбе занимљивије“ је у *BOW* приступу изгубио шири контекст и због термина „волео“ и „занимљивије“ означен као позитиван сентимент, уместо негативан. Приближан однос погрешно негативних анотација је било означено негативним сентиментом због *BOW* приступа.

Анализом грешака интеграционог модела групе модела дубоког учења за задатак идентификације сентимента (Табела XXVIII) је утврђено да постоји знатно мање типова грешака уочених код интеграционог модела групе класичних модела. Захваљујући претходно обученим моделима дубоког учења на корпусима општег домена је надомешћен недостатак одређених речи и израза. Међутим, улазне секвенце *ABSA* система у овом истраживању могу бити веома кратке и стога се контекст не може довољно уочити. На пример, сегмент текста „да се на часовима излаже само важно“ исказује критику на одржавање наставе али из контекста то није очигледно и означено је као позитиван сентимент. Такође, двосмисленост утиче на грешке модела. На пример, сегмент текста „Асистент треба да се угледа на професора“ је двосмислена, и има негативан сентимент из угла асистента јер се упућује да треба да се угледа на професора. Међутим, из угла професора је то комплимент и носи позитиван сентимент. Други тип грешака су везане за домен. На пример, сегмент текста „све ради асистент“ је у високом образовању означена као негативан сентимент јер указује да наставничка задужења ради асистент. Док из општег домена, исказан је позитиван сентимент јер може указивати да асистент ради све тј. да је вредан.

6.4 *Дискусија импликације и ограничења овог истраживања*

Циљ овог истраживања је био развој методологије за аспектно базирану сентимент анализу на нивоу сегмента реченице за домен високог образовања. Постоји велики број постојећих *ABSA* приступа прилагођених домену високог образовања. Међутим, дати приступи раде на нивоу документа (Chauhan *et al.,* 2018; Valakunde and Patwardhan, 2013) или реченице (Shaikh and Doudpotta, 2019; Sindhu *et al.,* 2019) што може бити проблематично ако су потребне детаљније информације о аспектима и додељеним сентимент поларитета.

Експериментални резултати ове студије показују да *ABSA* за домен високог образовања за српски језик може бити успешно примењен на нивоу сегмента реченице. Добијени резултати модела су у складу са последњим актуелним истраживањима у *ABSA* за домен високог образовања (Shaikh and Doudpotta, 2019; Sindhu *et al.,* 2019). Аутори (Shaikh and Doudpotta, 2019) су за задатак идентификације аспеката саопштили просечну прецизност од 0,83 и просечан одзив од 0,80, док је за задатак идентификације сентимента саопштена просечна тачност од 0,90. Аутори (Sindhu *et al.*, 2019) су саопштили просечну Ф-меру од 0,85 за идентификацију аспеката и просечну Ф-меру од 0,86 за идентификацију сентимент поларитета. Међутим, последња актуелна истраживања у домену високог образовања су спроведена на нивоу документа или

реченице. Новија *ABSA* истраживања у другим доменима попут домена рецензија ресторана (Wang *et al.*, 2019; Sun *et al.*, 2019) су заснована на моделима дубоког учења на нивоу реченица ради боље детекције контекста. Добијене Ф-мере крајњих модела система овог истраживања су сличне Ф-мерама поменутих истраживања. Аутори (Wang *et al.*, 2019) су за задатак идентификације аспекта постигли просечну Ф-меру од 0,87, док су за задатак идентификације сентимента постигли просечну Ф-меру од 0,85. Док су аутори (Sun *et al.*, 2019) остварили просечну Ф-меру од 0,88 за задатак идентификације аспеката, а за задатак идентификације сентимента су саопштили меру тачности од 0,93.

6.4.1 Разлике у перформансама на корпусима К1 и К2

На основу резултата модела се може приметити да постоји разлика између корпуса К1 и К2. Прва разлика је извор података односно окружење у ком су сачињени коментари. Корпус К1 је сачињен од коментара студената о наставном особљу и студијским програмима ФТН-а добијених из званичних студентских анкета спроведених на ФТН-у. Анкете су попуњаване ручно током семестра а потом су ручно преписане у базу података од стране административног особља факултета. На основу овог се може закључити да су студенти били у окружењу које намеће коришћење званичног језика при писању коментара. Дужина коментара варира од 2 до 80 речи, а нешто краћи коментари (3 до 15 речи) чине 74 процената корпуса. Ови подаци могу указивати на то да су коментари писани у ограничено време током предавања у већини краћи коментари. Корпус К2 представља колекцију јавно доступних коментара наставног особља разних факултета у Србији са веб сајта “Оцени професора”. За разлику од корпуса К1, рецензије корпуса К2 су постављане на Интернет слободном вољом корисника веб сајта за одабране професоре било ког факултета у Републици Србији. На основу овог се може закључити да су корисници, под претпоставком студенти, били у окружењу које не намеће коришћење званичног језика при писању коментара. Анализом обучавајућег скупа је утврђено постојање жаргонских и сентимент израза који се нису појављивали у корпусу К1. У резултатима појединачних модела (Табела XXV) се могло видети да сентимент речници, који су формирани на основу корпуса К2, не доприносе толиком побољшању перформанси појединачних модела на корпусу К1 као на корпусу К2 (нпр. просечна Ф-мера *DB* модела на К1 је 0,57 док је 0,79 на К2).

Такође, по дефиницији веб сајта “Оцени професора” корисници се наводе на коментарисање само наставника али не и студијских програма, што се може видети и по броју анотација корпуса К2 (Табела XIV). Време за писање коментара није било ограничено што је, може се приметити, утицало да дужина коментара варира од 1 до

180 речи, а да коментари средње дужине (6 до 30 речи) чине 53 процената корпуса. Такође, анализом грешака је утврђено да је постојала мала разлика приликом анотације корпуса К1 и К2 која је довела до мањег пада перформанси модела. Анотатори корпуса К2 нису били ограничени експлицитним анотирањем сегмената реченица (фраза и клауза), те стога анотације нису морале да се подударају са границама фраза и клауза. Те је због начина формирања скупова за обучавање, валидацију и тестирање довело до тога да издвојени сегмент реченице промени аспект или сентимент поларитет (Секција 6.3). На основу свега наведеног се може закључити да корпуси К1 и К2, иако су из истог домена, имају одређен степен различитости.

6.4.2 Разлике класичних модела и модела дубоког учења

На основу резултата модела се могу закључити разлике између групе класичних модела и групе модела дубоког учења. Модели групе класичних модела припадају зачетницима анализе у *NLP* области. Модели засновани на речницима и правилима веома зависе од анотираних ресурса који су често зависни од домена и конкретног корпуса. На пример, модел заснован на правилима који је развијен за корпус К1 је допринео *SVM* моделу побољшање перформанси за два процента док је за исти модел за корпус К2 допринео побољшању перформанси за један проценат (Табела XX). Такође, модел заснован на правилима је развијен за корпус К2 и остварио је просечну Ф-меру од 0,79 на корпусу К2 док је на корпусу К1 остварио за двадесет два процента мању просечну Ф-меру (0,57). Класични модели машинског учења се ослањају на ручно формиране језичке особине најчешће применом *BOW* модела. Током анализе грешака је утврђено да велики проценат грешака појединачних класичних модела машинског учења чине управо грешке због недостатка речи у речник термина и губитка контекста. У оквиру овог истраживања, за задатак идентификације аспеката је интеграциони модел из групе класичних модела остварио веома добре резултате, просечну Ф-меру од 0,81 за корпус К1 и просечну Ф-меру од 0,71 за корпус К2 (Табела XX).

Модели групе модела дубоког учења су тренутно најактуелнији приступи анализе у *NLP* области. Захваљујући комплексној репрезентацији језика и техници трансфера знања су модели дубоког учења остварили веома добре резултете у *NLP* задацима. У оквиру овог истраживања, захваљујући претходно обученим векторским репрезентацијама речи су у великој мери отклоњени проблеми недостатка речи у речнику модела и губитка контекста. Постигнуте су веома добре перформансе појединачног и интеграционог модела дубоког учења за задатак идентификације сентимента, просечна Ф-мера од 0,96 за корпус К1 и просечна Ф-мера од 0,90 за корпус

К2 (Табела XXVIII). Међутим, за задатак идентификације аспеката је интеграциони модел групе модела дубоког учења остварио за пет процената мању просечну Ф-меру од интеграционог модела групе класичних модела (Табела XXI, Табела XXIII). На основу анализе грешака је утврђено постојање конфузије модела за аспекте *Организација*, *Настава* и *Предмет*, која се може приписати кратким секвенцама текста из којих није препознат контекст. Такође, узрок може бити и недостатак већег броја анотираних примера на основу којих би се могуће конфузије разрешиле дообучавањем модела.

6.4.3 Примењивост предложеног система и ограничења истраживања

Једна од битних импликација овог истраживања јесте примењивост предложеног система у другим високошколским установама у Србији. Истраживање јавног мњења великих размера из анкета студената захтева аутоматизацију процеса анализе мишљења студената на основу ког се може побољшати квалитет процеса доношења одлука. Стога, предност и потреба предложеног система је вишеструка за високошколске установе.

Методологија која је предложена у оквиру овог истраживања се може применити у потпуности и на другим високошколским установама у Србији. Према потреби се може изменити шема анотације аспеката и приступити процесу анотације корпуса златног стандарда. Формирани систем у оквиру овог истраживања се у иницијалној фази може применити и на другим високошколским установама у Србији. Ради веће ефикасности система (нпр. за задатак идентификације аспеката) је неопходно дообучити коришћене моделе додатним корпусом из домена високог образовања. Примена формираних модела заснованих на правилима и речницима се показала мање успешном на другим корпусима због зависности од корпуса на којима су настали. Стога је за примену ових модела неопходно прилагођавање, ондосно допуна правила и речника новим знањем.

У наставку су изложена ограничења овог истраживања. Иако *ABSA* методологија може бити разматрана као језички независна, развијене компоненте модела машиснког учења, модела заснованог на речницима и правилима, примењива је једино за рецензије написаних на српском језику. Међутим, последње актуелни модели дубоког учења са претходно обученим језичким моделом на великим корпусима су значајно проширили примену модела на разне *NLP* задатке и језике. Тачније, модели дубоког учења коришћени у овом истраживању (*BERT*, *XLM*, *XLM-RoBERTa*) су вишејезични и обучени су на подацима за више од 100 језика. Стога, ови модели се могу применити и на другим језицима и доменима уз претходно дообучавање модела.

Као што је приказано експерименталним резултатима, ресурси као што су правила и речници сачињени на основу једног корпуса су осетљиви на врсту односно жаргон језика коришћен у рецензији. Више формални жаргон је коришћен у корпусу К1 у односу на мање формални у корпусу К2. Стога, један од ограничавајућих фактора за успешну примену ових ресурса је захтев за ручним финим подешавањем. На пример, коришћење сентимент речника за детекцију сентимента само одређеног скупа аспеката (Секција 6.2). На крају, техника сегментације реченице која је представљена у овом истраживању делимично се ослања на одређене врсте речи. Коришћени су везници, предлози и глаголски прилози прикупљених из корпуса К1 који можда нису присутни у сваком корпусу. Применом неког језичког алата попут РЕЛДИ алата за српски језик би побољшао ефикасност сегментације реченица. Такође, услед мањка података класа аспеката претходно обучени модели дубоког учења нису довољно дообучени за задатак идентификације аспеката. Стога је побољшање квалитета и квантитета корпуса кључно за дообучавање вишејезичне модела дубоког учења попут *BERT*, *XLM* и *XLM-RoBERTa* модела.

7 Могућности примене система за аутоматизовану анализу мишљења

Систем за аутоматску анализу мишљења развијеног у оквиру ове докторске дисертације има широке могућности примене у свим високошколским установама у Србији. Примена се пре свега огледа у аутоматизованом формирању извештаја који приказују аспекте студирања са којима су студенти задовољни или незадовољни. На основу формираних извештаја управа одређене високошколске установе може имати помоћ при доношењу одлука у циљу побољшања квалитета наставе и других аспеката везаних за студирање. Неки од примера извештаја дати су у наставку и засновани су на сродном истраживању (Grljević, 2016).

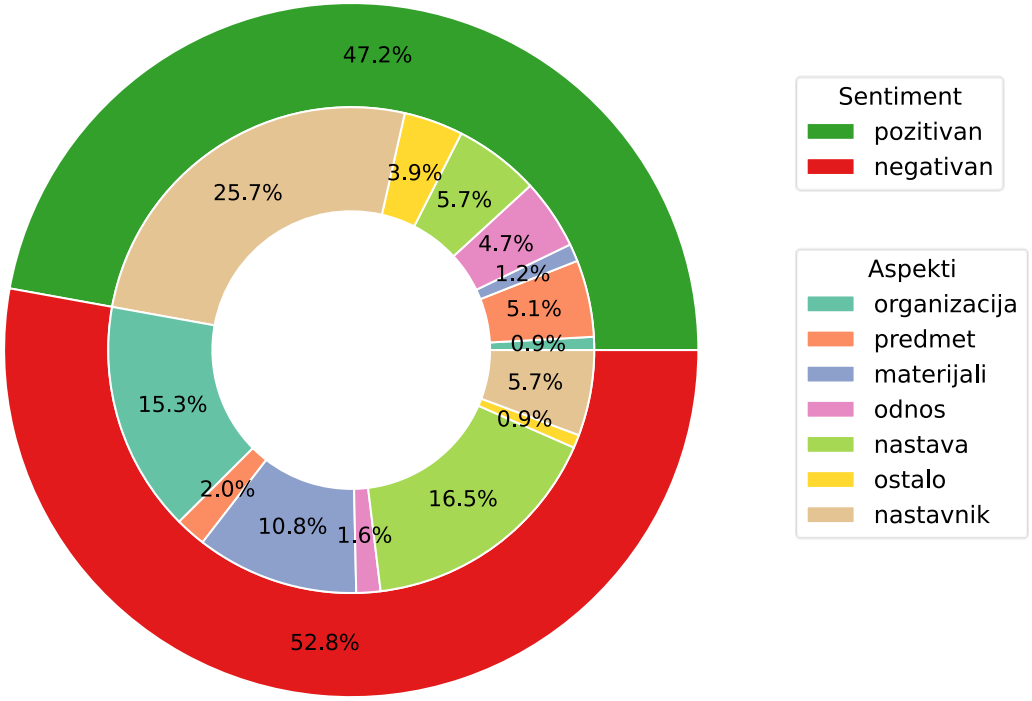
У зависности од типа, студентске анкете углавном садрже опште податке о анкетираној особи попут смера студирања, године уписа, године студирања, анкетираном предмету итд. На основу анализе мишљења из коментара анкета и ових општих података, могу се формирати извештаји који наводе следеће:

- да ли су анкетирани студенти опште задовољни или нису задовољни,
- којим аспектима студирања су студенти највише задовољни, односно нису задовољни,
- којим аспектима студирања су студенти највише задовољни, односно нису задовољни, на нивоу одређеног смера, предмета, године студирања, године уписа, или наставника,
- промена задовољства, односно незадовољства, током анкетираних година за одређене аспекте високог образовања за одређени смер, предмет, наставника или годину студија.

Поред општих података резултатима анализе мишљења могу се придружити и други подаци као што су: оцене предмета, пролазности на предмету, просечне оцене на предмету итд. У наставку су приказани типови графичких извештаја са додатним текстуалним описом тумачења резултата. Подаци у извештајима су делимично измењени како би се сачувала приватност власника података.

7.1 *Сумарни извештаји без временске компоненте*

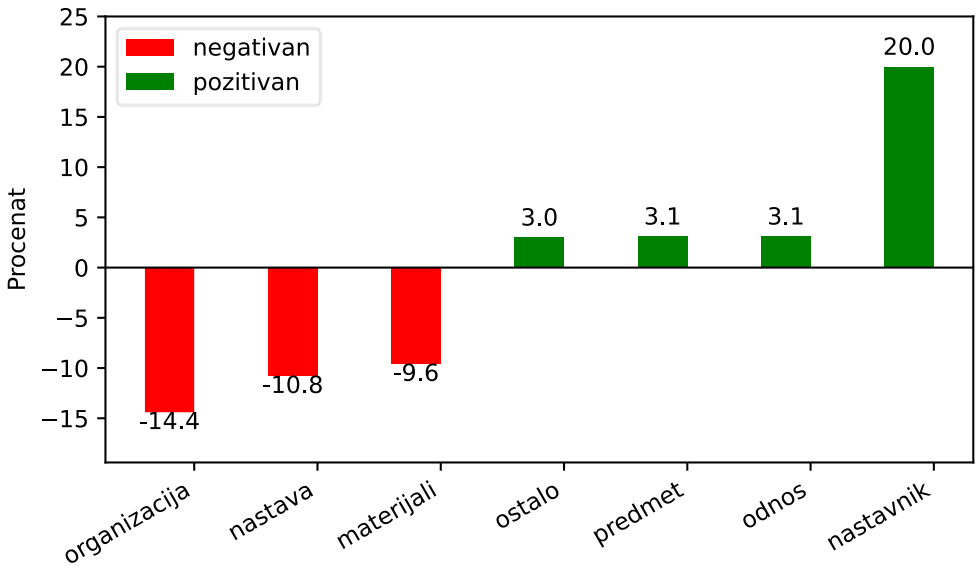
На Слика 57 је приказан графички извештај *ABSA* анализе на нивоу факултета.



Слика 57 *Приказ резултата ABSA анализе на нивоу целог факултета*

Статистички подаци приказани на Слика 57 показују да већина тј. 52,8 процената анализираних коментара чине сегменти означени негативним поларитетом. Највећи проценат негативног сегмената је упућено ка аспекту *Настава* (16,5 процената), *Организација* (15,3 процента) и *Материјали* (10,8 процената). То значи да половину негативног сентимента (26,1 процената од укупно 52,8 процената) је усмерено на аспекте предмета (*Организација* и *Материјали*). Преосталих 47,2 процената сегмената коментара су означени позитивним сентиментом. Највећи проценат позитивног сентимента (36,1 процената од укупно 47,2 процената) је упућен ка аспектима наставника – *Наставник* (25,7 процената), *Настава* (5,7 процената) и *Однос* (4,7 процената). На основу статистичких података се може закључити да су студенти ФТН-а, у периоду од 2011. до 2016. године, највише били незадовољни аспектима везаним за предмете, односно највише задовољни аспектима везаним за наставнике.

Укупан сентимент поларитет по аспектима се може добити сумирањем процената позитивног сентимента са позитивним предзнаком и процената негативног сентимента са негативним предзнаком. Уколико је укупан сентимент поларитет по аспекту већи од нуле, значи да позитиван сентимент преовлађује. Обратно, уколико је укупан сентимент поларитет по аспекту мањи од нуле, негативан сентимент поларитет преовлађује. Укупно највећи позитиван сентимент по аспекту је усмерен на аспект *Наставник*, док је укупно највећи негативан сентимент по аспекту усмерен на аспект *Организација*. Приказ збирних резултата *ABSA* анализе за одређени студијски програм ФТН-а је приказан на Слика 58.



Слика 58 *Збирни приказ резултата ABSA анализе на нивоу целог факултета*

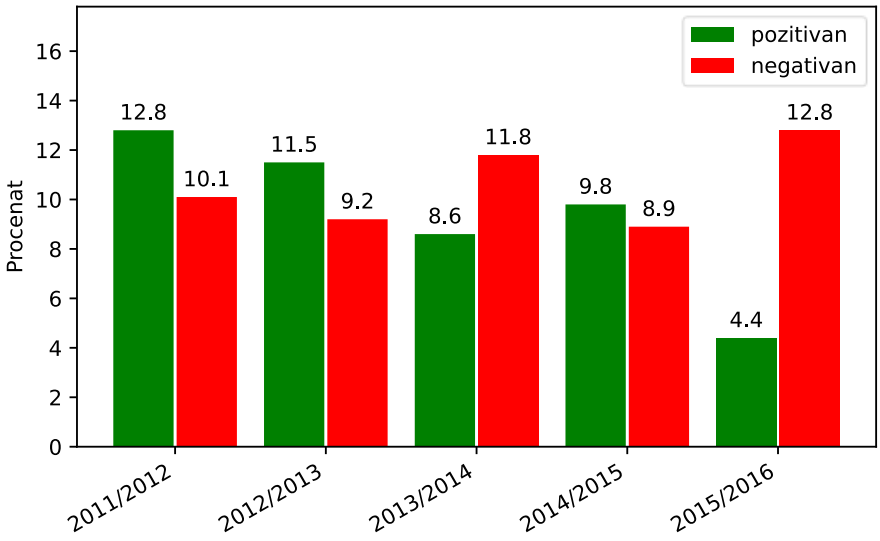
Генерални извештај о сентимент поларитету и аспектима високог образовања на нивоу факултета се може детаљније специфицирати према додатним подацима о студијском програму, предмету, наставнику и години студија. Овим специфичнијим типом извештаја се резултати *ABSA* анализе односе за одређени студијски програм, предмет или наставника, чиме се одређен проблем може свести на мању целину. На пример, на ФТН-у постоји преко 60 студијских програма, преко 3.000 предмета, преко 1.400 наставника, где генерални тип извештаја о сентименту и аспекту високог образовања можда не даје довољно детаљне информације. Овде је потребно извести *ABSA* анализу на мањој целини факултета. Графички прикази специфичних извештаја се формирају и тумаче на исти начин као и генерални извештаји приказани на Слика 57 и Слика 58.

7.2 *Извештаји са временском компонентом*

Извештаји са временском компонентом укључују генералне и специфичне типове извештаја са фокусом на време тј. године анкетирања. Циљ извештаја јесте праћење промене сентимент поларитета кроз време и утврђивање корелације са променом оцене предмета, оцене на предмету, пролазности на предмету итд.

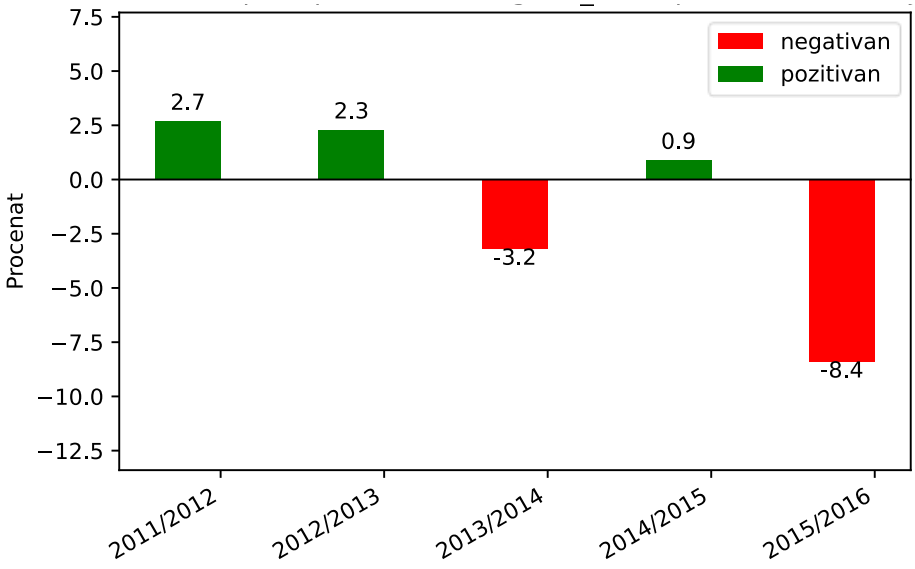
Извештаји са временском компонентом састоје се од четири типа графика. Први тип графика приказује промену сентимент поларитета по годинама. Како би се лакше утврдио преовлађујући сентимент поларитет по анкетној години употребљен је други тип графика који приказује промене сентимент поларитета кроз време. Трећи тип графика је фокусиран на резултате сентимент анализе у односу на аспекте високог образовања кроз време. За сваки период анкетирања (нпр. школска 2011/2012 година) је графички приказан процентуални однос позитивног и негативног сентимента у односу на анализиране аспекте. Проценти сентимент поларитета по аспектима на графику су приказани у односу на одређени период анкетирања. По угледу на други тип графика за преовлађујући сентимент поларитет по анкетној години је настао четврти тип графика - збирни приказ промена сентимент поларитета по аспектима за одређени анкетни период. Примери свих наведених типова извештаја дати су у наставку.

На Слика 59 је приказана промена сентимент поларитета по анкетним годинама. Као што се већ могло видети на генералном извештају (Слика 57), студенти су генерално на нивоу факултета у анкетном периоду више били незадовољни него задовољни. Задовољство је преовладавало у прве четири анкетне године, док се у школској школској 2015/2016 години смањио готово за трећину у односу на почетак посматраног периода. Незадовољство је варирало током посматраног периода и у школској 2015/2016 години порасло за петнаест процената у односу на почетак посматраног периода, школска 2011/2012 година.



Слика 59 *Промена сентимент поларитета кроз временску линију*

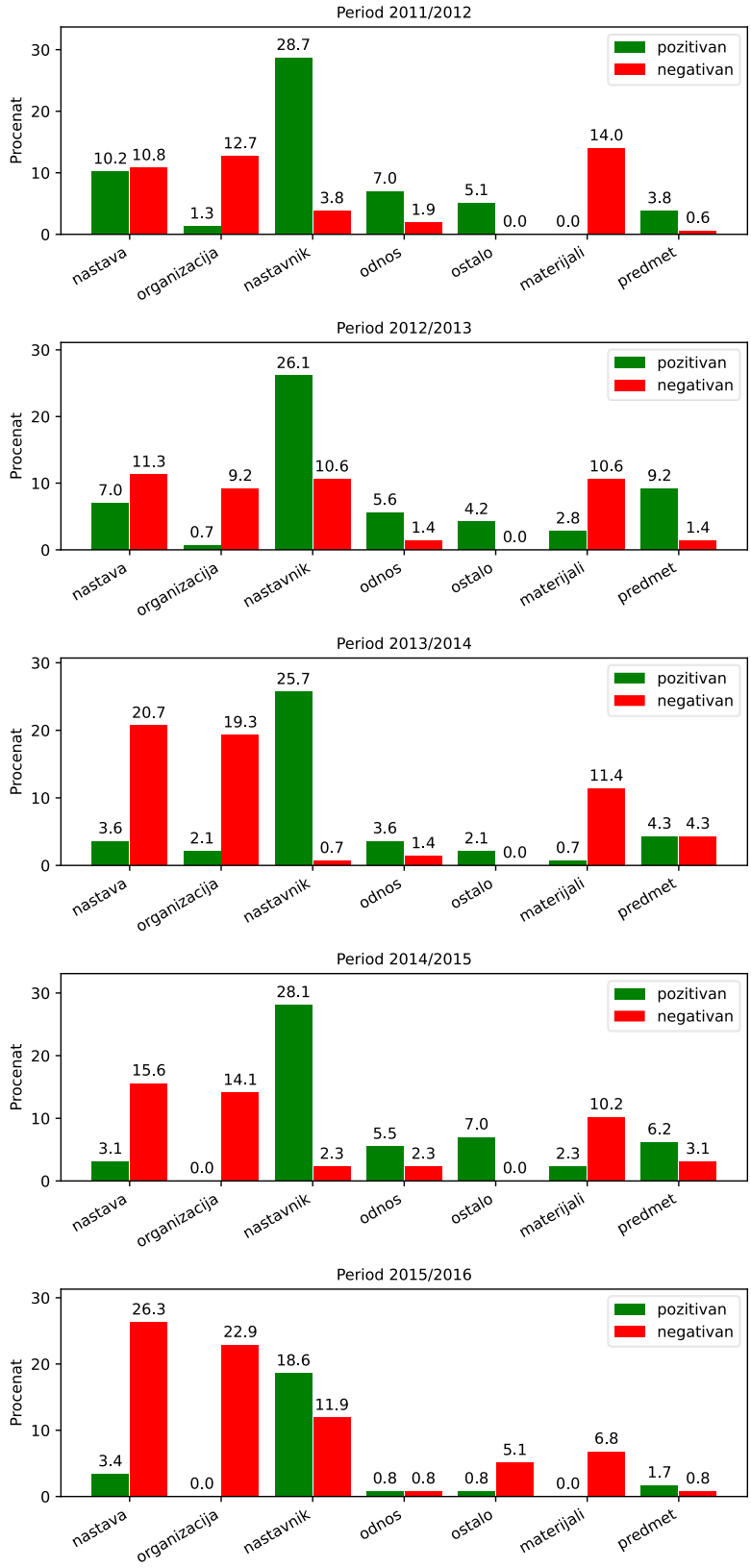
Укупна промена сентимент поларитета кроз анкетне године је приказана на Слика 60. Као што је већ уочено на претходном графику (Слика 59), студенти посматраног студијског програма су највише били задовољни у школској 2011/2012 години, а највише незадовољни у школској 2015/2016 години. Посматрајући засебно анкетне школске године, студенти су већином били задовољни у целом посматраном периоду анкетирања. Међутим, интензитет незадовољства у школској 2015/2016 години указује на потребу усмеравању додатне пажње на одређене аспекте.



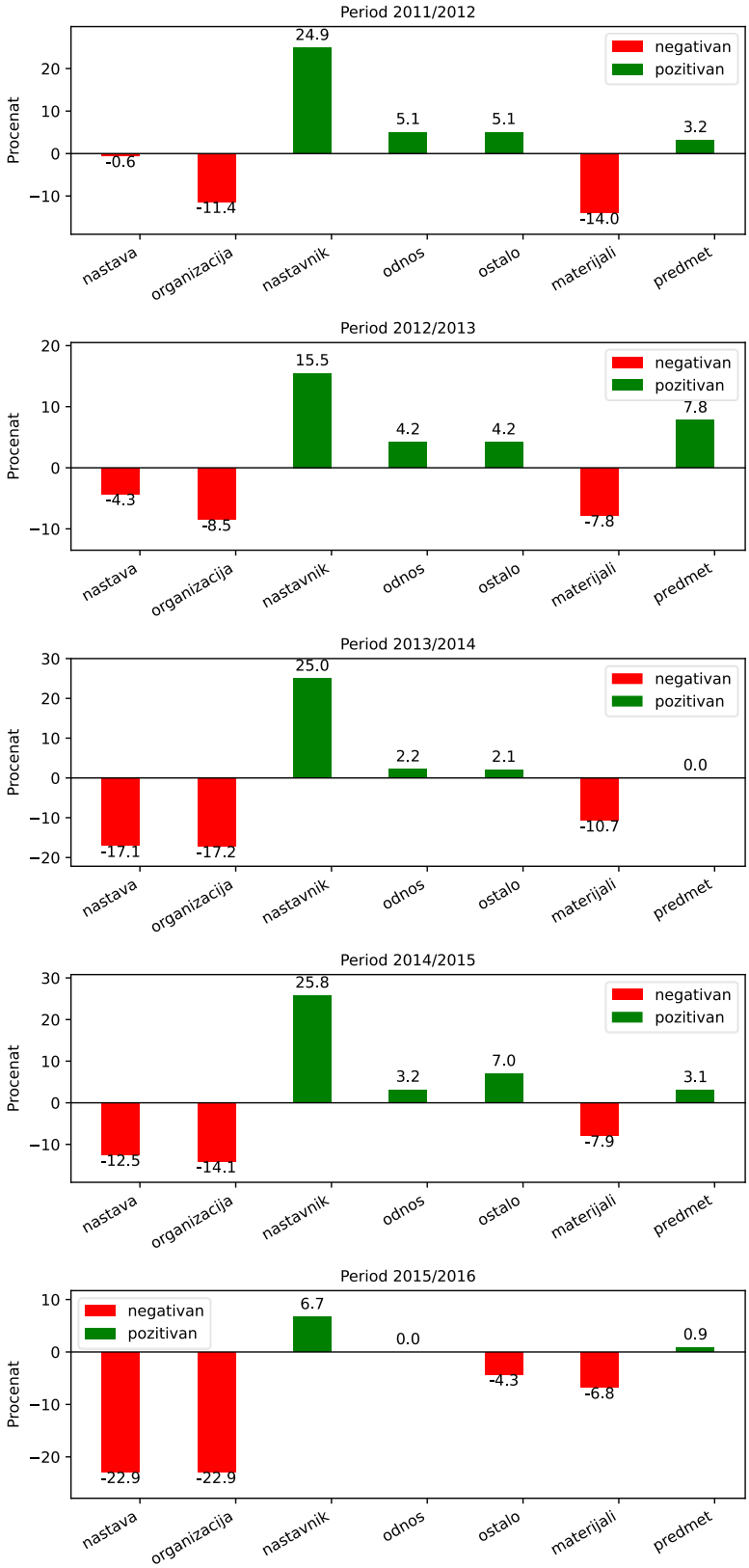
Слика 60 *Збирни приказ промене сентимент поларитета кроз временску линији*

Промена *ABSA* резултата по анкетним годинама је приказана на Слика 61, док је збирни приказ промене сентимента по аспектима приказан на Слика 62. Студенти су у посматраном периоду анкетирања од 2011. до 2016. године највише били незадовољни аспектима *Настава* и *Организација*, чији проценат негативног поларитета постепено повећавао од школске 2011/2012 године па све до школске 2015/2016 године, кад је установљено највеће незадовољство у посматраном периоду анкетирања. У посматраном периоду анкетирања, студенти су највише били задовољни аспектом *Наставник*, чији проценат позитивног сентимента је приближно исто висок у прве четири анкетне школске године. Након тога је у школској 2015/2016 години уследио пад од 30 процената у односу на просек претходне четири анкетне године (од школске 2011/2012 до 2014/2015 године). Задовољство студената аспектима *Предмет* и *Однос* се такође постепено смањивало у посматраном анкетном периоду.

Остали примери специфичнијих извештаја *ABSA* анализе за одређени студијски програм, предмет, наставника и годину студија, по анкетним годинама (од 2011. до 2016. године) се формирају и тумаче на исти начин.



Слика 61 *Промена ABSA резултата кроз временску линију*



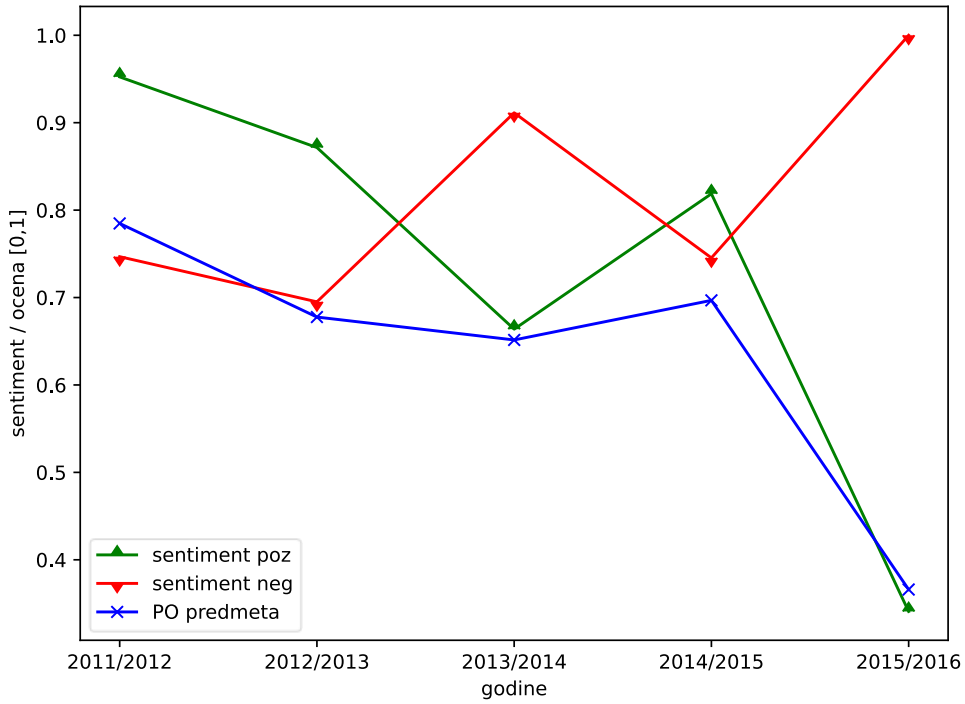
Слика 62 *Збирни приказ промена ABSA резултата кроз временску линију*

7.2.1 Извештаји који илуструју временске корелације

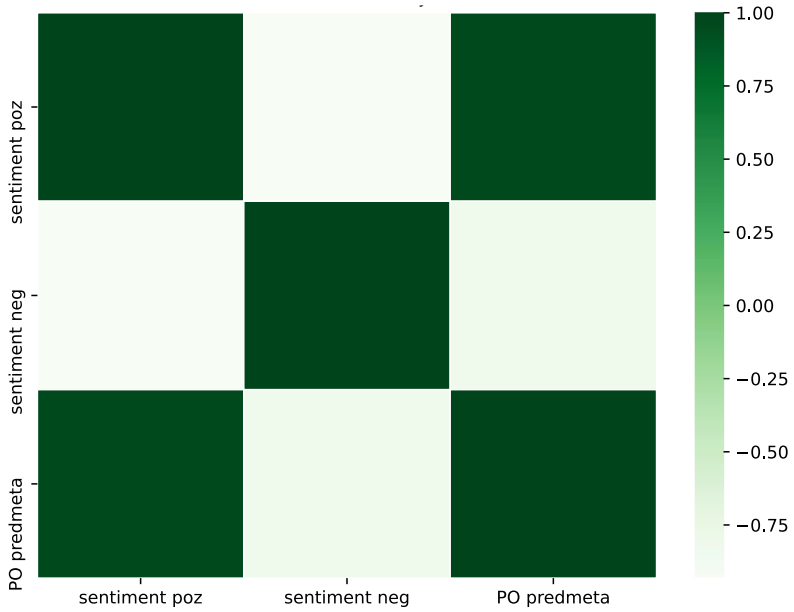
Извештаји (у облику графика) који укључују временску компоненту могу да се користе за утврђивање потенцијалне корелације промене сентимента и неког од релевантних података као на пример оцене предмета просечне оцене студената на предмету или пролазности на предмету. Први тип графика приказује промену сентимент поларитета кроз анкетне године. За лакше визуално утврђивање корелације је потребно све вредности које се пореде на графику сведене на опсег $[0, 1]$. Укупне вредности позитивног и негативног сентимент поларитета по анкетној школској години су подељене са укупним бројем анкета тј. коментара у датој школској години. Тиме се добио просечан број позитивног и негативног сентимент поларитета по анкетној школској години. Како би се осигурало да вредности припадају опсегу $[0, 1]$, добијене просечне вредности сентимент поларитета су подељене максималном вредношћу сентимент поларитета током целог посматраног анкетног периода. Оцене предмета и оцене на предмету су вредности од пет до десет, стога је свођење вредности на опсег $[0, 1]$ извршено применом минималне-максималне нормализације. Вредности су умањене за минималну вредност (у овом случају пет) и подељене разликом максималне и минималне вредности (у овом случају пет). Пролазност на предмету се добија дељењем бројем студената који су слушали предмет и бројем студената који су положили предмет. Тиме су добијене процентуалне вредности пролазности које су у опсегу $[0, 1]$. Други тип графика представља корелациону матрицу вредности са првог типа графика. Вредности близу један (или тамно зелена боја) означава велики степен корелације вредности. Док вредности мањих од нуле представљају непостојање корелације између посматраних података. Стога су вредности на главној дијагонали матрице јединице, јер се подаци пореде сами са собом.

Пример извештаја промене сентимента по годинама анкетирања у односу на просечну оцену предмета за одређени предмет ФТН-а је приказан на Слика 63. За утврђивање корелације података са графика, на Слика 64 је дата корелациона матрица. На Слика 63 се може видети висок степен корелације позитивног сентимента и оцене предмета, односно при високом проценту позитивног сентимент поларитета је оцена предмета висока. Падом позитивног сентимент поларитета и порастом негативног сентимент поларитета у школској 2013/2014 години је дошло и до пада оцене предмета. У наставку, школској 2014/2015 години, је дошло до пораста позитивног сентимент поларитета и пада негативног сентимент поларитета, што је узроковало порасту оцене предмета. У последњој анкетној школској години 2015/2016 је дошло до великог пораста негативног и пада позитивног сентимент поларитета што је узроковало и паду

просечне оцене предмета. Висока корелација позитивног сентимента и просечне оцене предмета се може видети и на корелационој матрици (Слика 64).

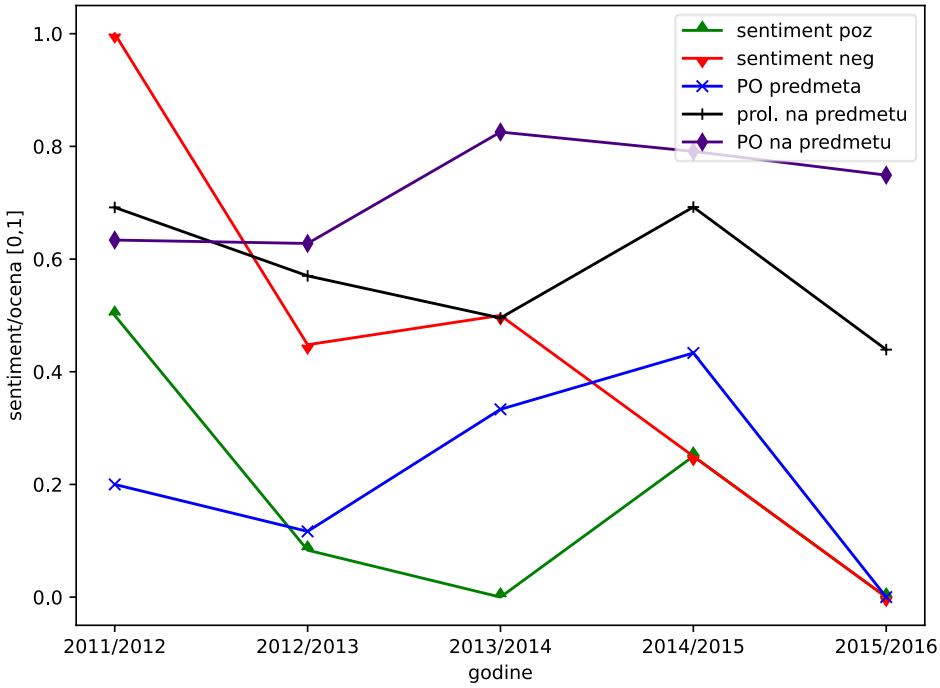


Слика 63 *Промена сентимент поларитета за одређени предмет кроз временску линију у односу на оцену предмета*

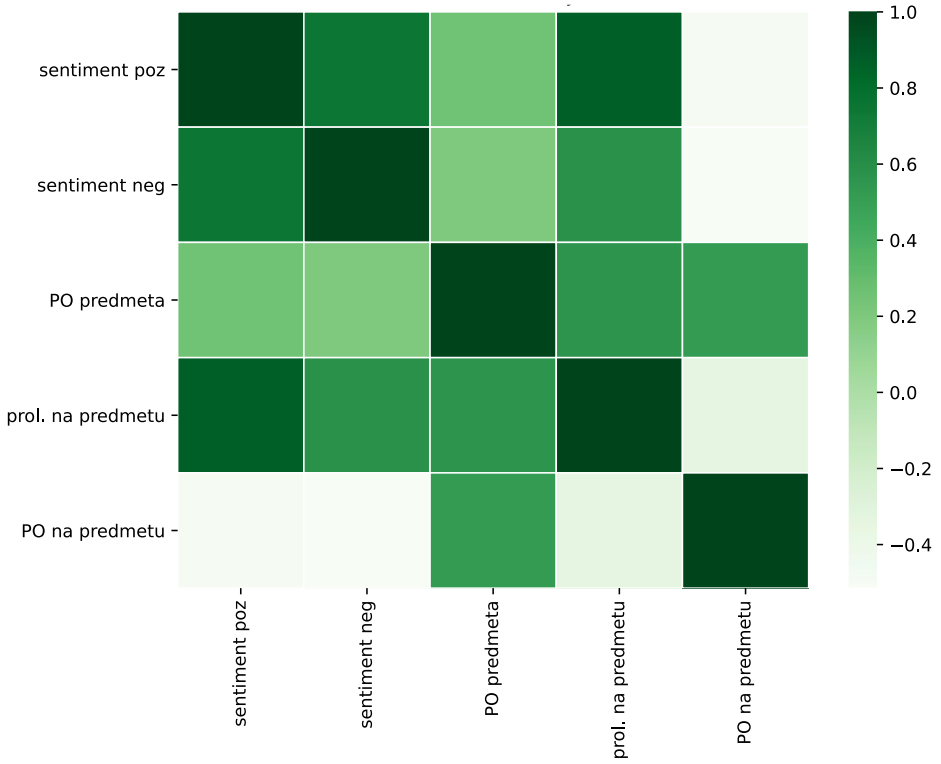


Слика 64 *Корелациона матрица промене сентимент поларитета за одређени предмет кроз временску линију у односу на оцену предмета*

Поред просечне оцене предмета, може се испитати корелација промена сентимента по годинама анкетирања за одређени предмет ФТН-а у односу на просечну оцену студената на предмету и пролазности студената на предмету. На Слика 65 је дат пример оваквог типа извештаја. У посматраном анкетном периоду се може видети постојање мањег степена зависности позитивног сентимент поларитета и пролазности студената на предмету. Промена просечне оцене анализираног предмета и просечне оцене на предмету није у корелацији са променом сентимента. Степен корелације се може утврдити анализом корелационе матрице (Слика 66) где је већи степен корелације забележен код података о пролазности студената на предмету.



Слика 65 *Промена сентимент поларитета за одређени предмет кроз временску линију у односу на оцену предмета, оцену на предмету и пролазности на предмету*



Слика 66 *Корелациона матрица промене сентимент поларитета за одређени предмет кроз временску линију у односу на оцену предмета, оцену на предмету и пролазности на предмету*

8 Закључак

Предмет истраживања ове докторске дисертације је формирање модела који омогућава аутоматизовано издвајање мишљења студената исказаних у текстуалним коментарима. Аутоматизовано издвајање мишљења припада делу научне под-области која се назива аспектно базирана сентимент анализа која је део шире области обраде природног језика. У разматрање се узимају коментари на српском језику исказани путем обавезних анкета које су спроведене на Факултету Техничких Наука (ФТН), Универзитета у Новом Саду. Методологија овог истраживања се базира на машинском учењу који такође укључује и компоненте модела заснованих на правилима и речницима. Предложена методологија је верификована имплементацијом прототипа система и евалуацијом истог. Такође, развијени модел је додатно евалуиран коришћењем корпуса из домена високог образовања који је резултат сродног истраживања. Извршена је компаративна анализа резултата система и изведени су закључци.

У првом поглављу је описан предмет истраживања докторске дисертације. Описана је потреба система за аутоматску анализу текстуалних коментара за институције виског образовања у Србији. Дефинисани су конкретни циљеви истраживања и хипотезе које су представљале оквир дисертације.

У другом поглављу су описане потребне теоријске основе везане за области које обухвата ова докторска дисертације. Дат је опис основних појмова обраде природног језика, сентимент анализе и аспектно базиране сентимент анализе. У наставку је описан општи методолошки оквир за развој система за обраду природних језика. Затим су описане теоријске основе приступа за аутоматизовану анализу мишљења из текстова, систематизовани по употребљеној методологији. Прво су дати приступи чија се методологија ослања на специјализоване речнике и ручно формирана правила. Након тога детаљно су приказани приступи чија се методологија базира на машинском учењу. Посебна пажња посевећена је методологијама које користе моделе дубоког учења јер они представљају тренутно најактуелније приступе у области обраде природног језика.

У трећем поглављу је дат преглед сродних релевантних истраживања. Фокус прегледа је пре свега на аспектно базираној сентимент анализи, али су поред тога обрађена и истраживања везана за сентимент анализу у домену високог образовања као и обраду природног језика на српском језику. У самом прегледу акценат је стављен на типове коришћених модела, ниво анализе (документ, реченица, сегмент реченице) и постигнуте резултате евалуације модела. Прво су описани првобитни приступи засновани на речницима и правилима, а затим приступи засновани на машинском учењу.

У четвртом поглављу описан је процес формирања корпуса златног стандарда који представља један од резултата ове докторске дисертације. Описан је процес анотације корпуса атрибутима потребним за извођење аспектно базиране сентимент анализе. У наставку су дати резултати анализе поузданости анотатора уз све потребне статистичке карактеристике корпуса. Упоредно је описан и корпус коментара из домена високог образовања који је резултат сродног истраживања, а употребљен је за додатну евалуацију модела развијеног у овој дисертацији.

У петом поглављу детаљно је приказан систем за аутоматску анализу мишљења развијеног у оквиру ове докторске дисертације. Архитектура система је заснована на модулима од којих сваки обавља значајан део процеса аспектно базиране сентимент анализе: пред-процесирање текста, издвајање важних језичких особина из текста, аутоматизовано одређивање аспеката (предмета мишљења) и аутоматизована додела сентимента (позитиван или негативан) сваком од аспеката. Развоју сваког од модула претходили су екстенизвни експерименти који су такође детаљно приказани у оквиру овог поглавља.

У шестом поглављу су описани експериментални резултати и дискусија система развијеног у оквиру дисретације. У оквиру дискусије прво су образложени резултати система као и резултати свих модела са којима се експериментисало у току развоја. Након тога дата су одговарајућа поређења са резултатима сродних истраживања, а дускутоване су и разлике између два корпуса који су употребљени за евалуацију. Веома значајан допринос овог дела шестог поглавља су запажања формирана након анализе грешака система које чине добру основу за будћа истраживања.

Свеобухватни резултати показују да се сентимент поларитет може успешно идентификовати у званичним студентским анкетама и рецензијама са интернета. Међутим, успешна идентификација аспеката на нивоу сегмента реченице зависи од различитих фактора, као што је величина корпуса, подударање сегмената реченице са анотацијама у обучавајућем скупу, лексичке варијабилности и фреквенције у обучавајућем скупу. Резултати такође демонстрирају да лексичка правила (без више апстрактних језичких особина као што су *POS* тагови или синтактичке зависности) могу

омогућити значајно побољшање перформанси само за корпус на основу ког су и развијени. Међутим, пажљивом интеграцијом модела заснованом на речницима са класичним моделом машинског учења може бити од користи без обзира на корпус.

Вишејезични модели дубоког учења са применом трансфера знања су се показали веома ефикасним у задатку идентификације сентимента. Претходно обучавање модела већим корпусима је омогућило повећање речника модела и препознавање корелације сентимент термина. У случају идентификације аспеката, модели дубоког учења су показали сличну успешност као и класични модели машинског учења. За побољшање перформанси овог модела је потребна већа количина анотираног корпуса за све класе аспеката којим би се квалитетније дообучавали модели дубоког учења. Међутим, важно је напоменути да су модели који су приказани у овом истраживању обучени на корпусу српског језика па је њихова употреба ограничена само на анкете писане на српском језику. Такође, специјализовани ресурси као што су речници и правилиа могу остварити лошије перформансе када се примењују на различите корпусе или могу захтевати прецизније подешавање.

На крају шестог поглавља је извршена дискусија ограничења система и потенцијала његове примењивости у другим институцијама виског образовања у Србији. Аутоматизацијом процеса анализе мишљења студената умањује се време и труд које је потребно људима задуженим за унос и обраду података из студентских анкета. Надгледањем аспеката којим су студенти задовољни или нису задовољни, администрација високошколске установе може лакше и брже доносити одлуке у циљу константног побољшања квалитета студирања. Такође, резултати овог истраживања могу помоћи у регрутовању нових и задржавању постојећих студената.

Предложена методологија се може применити у потпуности и на другим високошколским установама у Србији. Међутим, развијени прототип система односно компоненте модела машиснког учења, модела заснованог на речницима и правилима, су примењиве једино за рецензије написаних на српском језику. Развијени систем се може иницијално применити на другим високошколским установама у Србији али је, ради веће ефикасности система, неопходно дообучити коришћене моделе додатним корпусом из домена високог образовања. Последње актуелни модели дубоког учења са претходно обученим вишејезичким моделом на великим корпусима су значајно проширили примену модела на разне *NLP* задатке и језике. Стога, се развијени прототип модела може применити и на другим језицима и доменима уз претходно дообучавање модела.

У седмом поглављу су илустороване неке од могућности примене резултата система за аутоматизовану анализу мишљења из студентских анкета. Представљени су различити типови извештаја које институције виског образовања могу употребити у

процесу доношења одлука са циљем повећања квалитета студирања. Поред сумарних извештаја, приказани су и извештаји који укључују временску компоненту помоћу које се могу утврдити потенцијалне корелације мишљења студената са подацима као што су просечна оцена на предмету, пролазност на предмету итд.

Будући рад овог истраживања ће бити фокусиран на анализи различитих метода идентификације аспеката и сентимента на нивоу сегмента реченица. Испитивањем релација аспект-сентимент би се могло утврдити узрок задовољства или незадовољства аутора рецензије. Тиме би решење уоченог незадовољства било на решавању узрока негативног сентимента. Са друге стране, уочавањем узрока задовољства би допринело анализи шаблона који доводе до задовољства корисника. Део будућег рада ће бити усмерен и на побољшању квалитета и квантитета анотираног корпуса којим би се дообучавали модели дубоког учења и тиме очекивано остварили још боље резултате на задатку идентификације аспеката. Такође, један од задатака будућег рада може бити усмерен ка комбинацији *ABSA* резултата са техникама анализе података. На основу овог проширења би извештаји предложеног система дали још опсежније информације и предвиђања.

Литература

Alghunaim, A., Mohtarami, M., Cyphers, S. and Glass, J. (2015), “A Vector Space Approach for Aspect Based Sentiment Analysis”, *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, available at: https://doi.org/10.3115/v1/w15-1516 (Accessed: 23 October 2019).

Al-Smadi, M., Qawasmeh, O., Al-Ayyoub, M., Jararweh, Y. and Gupta, B. (2018), “Deep Recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels’ reviews”, *Journal of Computational Science*, Vol. 27, pp. 386–393.

Àlvarez-López, T., Juncal-Martínez, J., Fernández-Gavilanes, M., Costa-Montenegro, E. and González-Castaño, F.J. (2016), “GTI at SemEval-2016 Task 5: SVM and CRF for Aspect Detection and Unsupervised Aspect-Based Sentiment Analysis”, *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, available at: https://doi.org/10.18653/v1/s16-1049 (Accessed: 23 October 2019).

Batanovic, V. and Nikolic, B. (2017), “Sentiment classification of documents in Serbian: The effects of morphological normalization and word embeddings”, Telfor Journal, Vol. 9 No. 2, pp. 104–109.

Barnes, J., Lambert, P. and Badia, T. (2018), "Multibooked: A corpus of basque and catalan hotel reviews annotated for aspect-level sentiment classification", arXiv preprint arXiv:1803.08614.

Bhatnagar, V., Goyal, M. and Hussain, M.A. (2018), “A Novel Aspect Based Framework for Tourism Sector with Improvised Aspect and Opinion Mining Algorithm”, *International Journal of Rough Sets and Data Analysis*, Vol. 5 No. 2, pp. 119–130.

Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. (2017), "Enriching word vectors with subword information", Transactions of the Association for Computational Linguistics, Vol. 5 No. 1 , pp. 135-146

Boucher, J. and Osgood, C.E. (1969), “The pollyanna hypothesis”, *Journal of Verbal Learning and Verbal Behavior*, Vol. 8 No. 1, pp. 1–8.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S. (2020), "Language models are few-shot learners", arXiv:2005.14165.

Chauhan, G.S., Agrawal, P. and Meena, Y.K. (2018), "Aspect-Based Sentiment Analysis of Students' Feedback to Improve Teaching–Learning Process", *Information and Communication Technology for Intelligent Systems*, pp. 259–266.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y. (2014), "Learning phrase representations using RNN encoder-decoder for statistical machine translation", arXiv:1406.1078.

Cliche, M. (2017), "BB_twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs", ArXiv.org, available at: https://arxiv.org/abs/1704.06125 (accessed 24 January 2021).

Cohen, J. (1960), "A Coefficient of Agreement for Nominal Scales", *Educational and Psychological Measurement*, Vol. 20 No. 1, pp. 37–46.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. and Stoyanov, V. (2019), "Unsupervised cross-lingual representation learning at scale", arXiv:1911.02116.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V. and Salakhutdinov, R. (2019), "Transformer-xl: Attentive language models beyond a fixed-length context", ArXiv.Org, available at: https://arxiv.org/abs/1901.02860 (accessed 1 March 2020).

Danielsson, P.E. (1980), "Euclidean distance mapping", Computer Graphics and image processing, Vol. 14 No. 3, pp. 227-248.

Dave, K., Lawrence, S. and Pennock, D.M. (2003), "Mining the peanut gallery", *Proceedings of the Twelfth International Conference on World Wide Web - WWW '03*, available at: https://doi.org/10.1145/775152.775226 (Accessed: 23 October 2019).

Devlin, J., Chang, M. W., Lee, K. and Toutanova, K. (2018), "Bert: Pre-training of deep bidirectional transformers for language understanding", ArXiv.Org, available at: https://arxiv.org/abs/1810.04805 (accessed 1 March 2020).

Dohaiha, H.H., Prasad, P., Maag, A. and Alsadoon, A. (2019), "Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review", *Expert Systems with Applications*, Vol. 118, pp. 272–299.

Fleiss, J.L. (1971), "Measuring nominal scale agreement among many raters", *Psychological Bulletin*, Vol. 76 No. 5, pp. 378-382.

Fleiss, J.L., Levin, B. and Paik, M.C. (2003), “Statistical Methods for Rates and Proportions”, *Wiley Series in Probability and Statistics*, available at: https://doi.org/10.1002/0471445428 (Accessed: 23 October 2019).

García-Díaz, J. A., Cánovas-García, M., and Valencia-García, R. (2020), "Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in Latin America", Future Generation Computer Systems, Vol. 112 No. 1, pp. 641-657.

Goldberg, Y. (2017), "Neural network methods for natural language processing", Synthesis Lectures on Human Language Technologies, Vol. 10 No. 1, pp.1-309.

Grljević, O. (2016), Sentiment u sadržajima sa društvenih mreža kao instrument unapređenja poslovanja visokoškolskih institucija (Ph.D.), Univerzitet u Novom Sadu, Ekonomski fakultet u Subotici, Srbija.

Grljević, O., Bošnjak, Z. and Kovačević, A. (2020), “Opinion mining in higher education: a corpus-based approach”, Enterprise Information Systems, DOI: 10.1080/17517575.2020.1773542

Gupta, V., Singh, V.K., Mukhija, P. and Ghose, U. (2019), “Aspect-based sentiment analysis of mobile reviews”, *Journal of Intelligent and Fuzzy Systems*, Vol. 36 No. 5, pp. 4721–4730.

Hercig, T., Brychcín, T., Svoboda, L. and Konkol, M. (2016), “UWB at SemEval-2016 Task 5: Aspect Based Sentiment Analysis”, *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, available at: https://doi.org/10.18653/v1/s16-1055 (Accessed: 23 October 2019).

Hinton, G.E., Krizhevsky, A. and Wang, S.D. (2011), "Transforming auto-encoders", International conference on artificial neural networks, Vol. 1 No. 1, pp. 44-51.

Hogenboom, A., Bal, D., Frasincar, F., Bal, M., de Jong, F. and Kaymak, U. (2013), “Exploiting emoticons in sentiment analysis”, *Proceedings of the 28th Annual ACM Symposium on Applied Computing - SAC ’13*, available at: https://doi.org/10.1145/2480362.2480498 (Accessed: 23 October 2019).

Hochreiter, S. and Schmidhuber, J. (1997), "Long short-term memory", Neural computation, Vol. 9 No. 8, pp.1735-1780.

Howard, J. and Ruder, S. (2018), "Universal language model fine-tuning for text classification", ArXiv.Org, available at: https://arxiv.org/abs/1801.06146 (accessed 1 March 2020).

Hutchins, J. (1999), "Retrospect and Prospect in Computer-Based Translation", Proceedings of MT Summit VII, pp. 30–44.

Hu, M. and Liu, B. (2004), “Mining and summarizing customer reviews”, *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’04*, available at: https://doi.org/10.1145/1014052.1014073 (Accessed: 23 October 2019).

Ide, N. and Pustejovsky, J. eds. (2017), Handbook of linguistic annotation, Springer.

Jin, W., Ho, H.H. and Srihari, R.K. (2009), “OpinionMiner: A Novel Machine Learning System for Web Opinion Mining and Extraction”, *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’09*, available at: https://doi.org/10.1145/1557019.1557148 (Accessed: 23 October 2019).

Jurafsky, D. and Martin, J.H. (2009), *Speech and Language Processing (2nd Edition)*, Prentice-Hall, Inc., USA.

Kešelj, V. and Šipka, D. (2008), "A Suffix Subsumption-Based Approach to Building Stemmers and Lemmatizers for Highly Inflectional Languages with Sparse Resources", *INFOtheca - Journal of Informatics and Librarianship*, Vol. 9 No. 2, pp. 23a–33a.

Kim, S.M. and Hovy, E. (2007), “Crystal: Analyzing Predictive Opinions on the Web”, *ACL Anthology*, pp. 1056–1064.

Kiritchenko, S., Zhu, X., Cherry, C. and Mohammad, S. (2014), “NRC-Canada-2014: Detecting Aspects and Sentiment in Customer Reviews”, *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, available at: https://doi.org/10.3115/v1/s14-2076 (Accessed: 23 October 2019).

Kovačević, A., Grljević, O., Bošnjak, Z. and Svilengaćin, G. (2020), "The linguistic construction of sentiment expressions in student opinionated content: A corpus-based study", Poznan Studies in Contemporary Linguistics, 56(2), pp.207-249.

Kralj Novak, P., Smailović, J., Sluban, B. and Mozetič, I. (2015), “Sentiment of Emojis”, *PLOS ONE*, Vol. 10 No. 12, pp. e0144296.

Krippendorff, K. (2018), *Content Analysis*, SAGE Publications Inc, London, available at: https://us.sagepub.com/en-us/nam/content-analysis/book258450 (accessed 23 October 2019).

Krstev, C., Pavlovic-Lazetic, G., Vitas, D. and Obradovic, I. (2004), "Using textual and lexical resources in developing serbian wordnet", Romanian Journal of Information Science and Technology, Vol. 7 No. 2, pp. 147-161.

Lample, G. and Conneau, A. (2019), "Cross-lingual language model pretraining", arXiv:1901.07291.

Landis, R. and Koch, G. (1977), "The measurement of observer agreement for categorical data", Biometrics, pp.159-174.

Levitz, R.N. (2020), "2020 Cost of Recruiting an Undergraduate Student Report", available at: https://learn.ruffalonl.com/rs/395-EOG-977/images/2020_CostRecruiting_Report.pdf (accessed 26 October 2020).

Li, F., Han, C., Huang, M., Zhu, X., Xia, Y.J., Zhang, S. and Yu, H. (2010), "Structure-aware review mining and summarization", *Proceedings of the 23rd International Conference on Computational Linguistics, Association for Computational Linguistics*, pp. 653–661.

Li, S.C. (2013), "Exploring the Relationships among Service Quality, Customer Loyalty and Word-Of-Mouth for Private Higher Education in Taiwan", *Asia Pacific Management Review*, Vol. 18 No. 4, pp. 375–389.

Liu, B. (2015), *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*, Cambridge University Press, New York, available at: https://www.cambridge.org/rs/academic/subjects/computer-science/knowledge-management-databases-and-data-mining/sentiment-analysis-mining-opinions-sentiments-and-emotions?format=HB (accessed 23 October 2019).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019), "Roberta: A robustly optimized bert pretraining approach", arXiv:1907.11692.

Ljubešić, N., Boras, D. and Kubelka, O. (2007), "Retrieving Information in Croatian: Building a Simple and Efficient Rule-Based Stemmer", *1. međunarodna znanstvena konferencija "The Future of Information Sciences: INFuture2007 – Digital Information and Heritage"*, pp. 313-320.

Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S. and McClosky, D. (2014), "The Stanford CoreNLP natural language processing toolkit", Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pp. 55-60.

Milošević, N. (2012), "Stemmer for Serbian language", ArXiv.Org, available at: https://arxiv.org/abs/1209.4471 (accessed 23 October 2019).

Mladenović, M. (2016), "Informatički modeli u analizi osećanja zasnovani na jezičkim resursima", available at: https://doi.org/http://hdl.handle.net/123456789/4422 (accessed 23 October 2019).

Mladenović, M., Mitrović, J., Krstev, C. and Vitas, D. (2016), “Hybrid sentiment analysis framework for a morphologically rich language”, Journal of Intelligent Information Systems, Vol. 46 No. 3, pp. 599–620.

Mowlaei, M.E., Abadeh, M.S. and Keshavarz, H., (2020), "Aspect-based sentiment analysis using adaptive aspect-based lexicons", Expert Systems with Applications, Vol. 148 No. 1.

Norouzi, M., Fleet, D.J. and Salakhutdinov, R.R. (2012), "Hamming distance metric learning", Advances in neural information processing systems, pp. 1061-1069

Pang, B. and Lee, L. (2005), “Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales”, *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL ’05*, available at: https://doi.org/10.3115/1219840.1219855 (accessed 23 October 2019).

Pang, B. and Lee, L. (2008), “Opinion Mining and Sentiment Analysis”, *Foundations and Trends in Information Retrieval*, Vol. 2 No. 1–2, pp. 1–135.

Pang, B., Lee, L. and Vaithyanathan, S. (2002), “Thumbs up?: sentiment classification using machine learning techniques”, *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - EMNLP ’02*, available at: https://doi.org/10.3115/1118693.1118704 (accessed 23 October 2019).

Patwa, P., Aguilar, G., Kar, S., Pandey, S., PYKL, S., Gambäck, B., Chakraborty, T., et al. (2020), “SemEval-2020 Task 9: Overview of Sentiment Analysis of Code-Mixed Tweets”, ArXiv.org, available at: https://arxiv.org/abs/2008.04277 (accessed 1 October 2020).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. and Blondel, M. (2012), “Scikit-learn: Machine Learning in Python”, ArXiv.Org, available at: https://arxiv.org/abs/1201.0490 (accessed 23 October 2019).

Pennington, J., Socher, R. and Manning, C.D. (2014), "Glove: Global vectors for word representation", Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L. (2018), "Deep contextualized word representations", ArXiv.Org, available at: https://arxiv.org/abs/1802.05365 (accessed 1 March 2020).

Platt, J. (1999), "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods", Advances in large margin classifiers, Vol. 10, No. 3, pp.61-74.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M. and Al-Ayyoub, M. (2016), "SemEval-2016 Task 5: Aspect Based Sentiment Analysis", *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, available at: https://doi.org/10.18653/v1/s16-1002 (accessed 23 October 2019).

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S. and Androutsopoulos, I. (2015), "SemEval-2015 Task 12: Aspect Based Sentiment Analysis", *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, available at: https://doi.org/10.18653/v1/s15-2082 (accessed 23 October 2019).

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I. and Manandhar, S. (2014), "SemEval-2014 Task 4: Aspect Based Sentiment Analysis", *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, available at: https://doi.org/10.3115/v1/s14-2004 (accessed 23 October 2019).

Popescu, A.M. and Etzioni, O. (2007), "Extracting Product Features and Opinions from Reviews", *Natural Language Processing and Text Mining*, pp. 9–28.

Pregibon, D. (1981), "Logistic regression diagnostics", The Annals of Statistics, Vol. 9 No. 4, pp. 705-724.

Pustejovsky, J. and Stubbs, A. (2012), Natural Language Annotation for Machine Learning: A guide to corpus-building for applications, 1st ed., Vol. , O'Reilly Media, Inc.

Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I. (2018), "Improving language understanding by generative pre-training", available at: https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf (accessed 1 March 2020).

Rauschnabel, P.A., Krey, N., Babin, B.J. and Ivens, B.S. (2016), "Brand management in higher education: the university brand personality scale", *Journal of Business Research*, Vol. 69 No. 8, pp. 3077-3086.

Rosenthal, S., Farra, N. and Nakov, P. (2017), "SemEval-2017 Task 4: Sentiment Analysis in Twitter", ArXiv.org, available at: https://arxiv.org/abs/1912.00741 (23 October 2019).

Ruder, S., Ghaffari, P. and Breslin, J.G. (2016), "INSIGHT-1 at SemEval-2016 Task 5: Deep Learning for Multilingual Aspect-based Sentiment Analysis", ArXiv.Org, available at: https://arxiv.org/abs/1609.02748 (accessed 23 October 2019).

Saeidi, M., Bouchard, G., Liakata, M. and Riedel, S. (2016), "SentiHood: Targeted Aspect Based Sentiment Analysis Dataset for Urban Neighbourhoods", *ACL Anthology*, pp. 1546–1556.

Saias, J. (2015), "Sentiue: Target and aspect based sentiment analysis in semeval-2015 task 12", *Association for Computational Linguistics*, pp. 767-771.

Salton, G. and Mcgill, M.J. (1986), *Introduction to Modern Information Retrieval*, Mcgraw-Hill Intern, Auckland, available at: https://dl.acm.org/citation.cfm?id=576628 (accessed 23 October 2019).

Shaikh S. and Doudpotta, S.M. (2019), "Aspects Based Opinion Mining for Teacher and Course Evaluation", *Sukkur IBA Journal of Computing and Mathematical Sciences*, Vol. 3 No. 1, pp. 34–43.

Sim, J. and Wright, C.C. (2005), "The kappa statistic in reliability studies: use, interpretation, and sample size requirements", *Physical therapy*, Vol. 85 No. 3, pp. 257-268.

Sindhu, I., Daudpota, S.M., Badar, K., Bakhtyar, M., Baber, J. and Nurunnabi, M. (2019), "Aspect-Based Opinion Mining on Student's Feedback for Faculty Teaching Performance Evaluation", *IEEE Access*, Vol. 7, pp. 108729–108741.

Smeeton, N. (1985), "Early History of the Kappa Statistic", Biometrics, Vol. 41, No. 3, pp. 795-795.

Sun, C., Huang, L. and Qiu, X. (2019), "Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence", ArXiv.Org, avaiable at: https://arxiv.org/abs/1903.09588 (accessed 1 March 2020).

Taboada, M. (2016), "Sentiment analysis: an overview from linguistics", *Annual Review of Linguistics*, Vol. 2, pp. 325-347.

Tang, D. (2015), "Sentiment-Specific Representation Learning for Document-Level Sentiment Analysis", *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*, pp. 447-452.

Tang, D., Qin, B. and Liu, T. (2015), "Learning Semantic Representations of Users and Products for Document Level Sentiment Classification", *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pp. 1014-1023.

Thet, T.T., Na, J.C. and Khoo, C.S.G. (2010), "Aspect-based sentiment analysis of movie reviews on discussion boards", *Journal of Information Science*, Vol. 36 No. 6, pp. 823–848.

Turney, P.D. (2002), "Thumbs up or thumbs down?", *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, available at: https://doi.org/10.3115/1073083.1073153 (accessed 23 October 2019).

Valakunde, N.D. and Patwardhan, M.S. (2013), “Multi-aspect and Multi-class Based Document Sentiment Analysis of Educational Data Catering Accreditation Process”, *2013 International Conference on Cloud and Ubiquitous Computing and Emerging Technologies*, available at: https://doi.org/10.1109/cube.2013.42 (accessed 23 October 2019).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017), "Attention is all you need", In Advances in neural information processing systems, pp. 5998-6008.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. and Bowman, S.R. (2018), "GLUE: A multi-task benchmark and analysis platform for natural language understanding", arXiv preprint arXiv:1804.07461.

Wang, B. and Liu, M. (2015), "Deep learning for aspect-based sentiment analysis", Stanford University report, available at: https://cs224d.stanford.edu/reports/WangBo.pdf (accessed 23 October 2019).

Wang, H. and Castanon, J.A. (2015), “Sentiment expression via emoticons on social media”, *2015 IEEE International Conference on Big Data (Big Data)*, available at: https://doi.org/10.1109/bigdata.2015.7364034 (accessed 23 October 2019).

Wang, L. and Zhao, X. (2012), "Improved KNN classification algorithms research in text categorization", 2nd International Conference on Consumer Electronics, Communications and Networks, pp. 1848-1852.

Wang, Y., Sun, A., Huang, M. and Zhu, X. (2019), “Aspect-level Sentiment Analysis using AS-Capsules”, *The World Wide Web Conference on - WWW ’19*, available at: https://doi.org/10.1145/3308558.3313750 (accessed 23 October 2019).

Wiebe, J., Wilson, T. and Cardie, C. (2005), "Annotating expressions of opinions and emotions in language", Language resources and evaluation, Vol. 39 No. 2, pp. 165-210.

Wu, H., Gu, Y., Sun, S. and Gu, X. (2016), “Aspect-based Opinion Summarization with Convolutional Neural Networks”, *2016 International Joint Conference on Neural Networks (IJCNN)*, available at: https://doi.org/10.1109/ijcnn.2016.7727602 (accessed 23 October 2019).

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K. and Klingner, J. (2016), "Google's neural machine translation system: Bridging the gap between human and machine translation", arXiv preprint arXiv:1609.08144 (accessed 23 October 2019).

Xue, W. and Li, T. (2018), “Aspect Based Sentiment Analysis with Gated Convolutional Networks”, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, available at: https://doi.org/10.18653/v1/p18-1234 (accessed 23 October 2019).

Xu, J., Chen, D., Qiu, X. and Huang, X. (2016), “Cached Long Short-Term Memory Neural Networks for Document-Level Sentiment Classification”, ArXiv.Org, available at: https://arxiv.org/abs/1610.04989 (accessed 23 October 2019).

Yahya, A.H., Azizam, A.A. and Mazlan, D.B. (2014), "The impact of electronic words of mouth (eWOM) to the brand determination of higher education in Malaysia: From the perspective of middle east’s student", *Journal of Mass Communication Journalism*, Vol. 4, pp. 1-4.

Yang, S., Rosenfeld, J. and Makutonin, J. (2018), "Financial Aspect-Based Sentiment Analysis using Deep Representations", ArXiv.Org, available at: https://arxiv.org/abs/1808.07931 (accessed 1 March 2020).

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R. and Le, Q.V. (2019), "Xlnet: Generalized autoregressive pretraining for language understanding", In Advances in neural information processing systems, pp. 5753-5763.

Yi, J., Nasukawa, T., Bunescu, R. and Niblack, W. (2003), “Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques”, *Third IEEE International Conference on Data Mining*, available at: https://doi.org/10.1109/icdm.2003.1250949 (accessed 23 October 2019).

Kim, Y. (2014), “Convolutional Neural Networks for Sentence Classification”, ArXiv.org, available at: https://arxiv.org/abs/1408.5882 (accessed 6 February 2020).

Zhang, Y. and Wallace, B. (2015), "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification", arXiv preprint arXiv:1510.03820.

Биографија

Никола Николић је рођен 5.6.1988. године у Лозници, Република Србија. Одељење електротехничара рачунара Средње техничке школе у Лозници је завршио одличним успехом. Студије на Факултету техничких наука у Новом Саду је уписао 2007. године на студијском програму Рачунарство и аутоматика. Дипломирао је 2011. године са темом „Реализација *DLNA* медија плејер апликације за Андроид платформу“. Исте године је уписао мастер академске студије на студијском програму Рачунарство и аутоматика и одбранио мастер рад 2012. године са темом „Једно решење имплементације и интеграције *PMA* програмске подршке телевизијских пријемника“. По завршетку мастер студија 2012. године се запослио као сарадник у настави на Факултету техничких наука Универзитета у Новом Саду.

Школске 2013/2014 године је уписао докторске студије на Факултету техничких наука – смер Рачунарство и аутоматика, усмерење Примењене рачунарске науке и информатика. У јануару 2015. године је изабран је у звање асистента за ужу научну област Примењене рачунарске науке и информатика на истом факултету. Положио је све испите предвиђене планом и програмом студијског програма Рачунарство и аутоматика са просечном оценом десет. Аутор је девет публикованих научних радова. Учествовао је у изради више стручних и научних пројеката.

Живи у Новом Саду. Од страних језика говори енглески језик.

Прилози

A. Правила за идентификацију аспеката

У наставку су приложена правила која су коришћена у модулу за идентификацију аспеката. Правила су заснована на регуларним изразима којима се идентификују фразе које су коршћене у домену високог образовања.

Аспект	**Правила**
Наставник	[r'(\s|^)((asis|prof|nastavn)\w*)(\w+)?[ ](fenomenal|konfuz|\w*sigurn|supe|izuzet|dobar|los.?|\w*spor|zna|\w*razum|odlic.?n|monoton|interesant|najbolj|uplas.?en|izuzet|pozitiv|negativ|primer|rasejan|arogant|legenda|struc.?njak|profesional)\w*)', r'(\s|^)((fenomenal|konfuz|\w*sigurn|supe|izuzet|dobar|los.?|\w*brz|\w*spor|zna|\w*razum|odlic.?n|monoton|interesant|naj|uplas.?en|izuzet|pozitiv|negativ|primer|rasejan|arogant|legenda|struc.?njak|profesional)\w*)[ ](\w+)?((asis|prof|nastavn|c.?ovek|pedagog)\w*)', r'(\s|^)(jedan|jedna|jedni)[ ](od naj\w*)[ ](osoba|asis|prof|nastavni|pedagog|ljud)\w*)', r'(\s|^)(vis.?e)[ ](\w*)?(\w*)(kao s.?to je)', r'(\s|^)(je|se|su)[ ](z.?ivc.?an|trud|autoritet)\w*)', r'(\s|^)(je|su)[ ](\w*)?((struc.?nj)\w*)', r'([\n\r\s]|^)((svi|sve|svaka)[ ]pohval\w*)[ ](\w*)?(\w*)?((asis|prof|nastavn|njeg|nju|njih)\w*)', r'([\n\r\s]|^)((asis|prof|nastavn|njeg|nju|njih)\w*)[ ](\w*)?(\w*)?((svi|sve|svaka)[ ]pohval\w*)']
Настава	[r'(\s|^)((izlaga|izlaz.?)\w*)[ ](gradi|materij|predmet)\w*', r'(\s|^)((glasn|spor|brz|\w*zanimljiv|\w*razum|divno|\w*jasn|\w*precizn|\w*konzisten)\w*)[ ](\w*)?((izlaga|izlaz.?)\w*)', r'(\s|^)(nac.?i\w*)[ ](izvodje|predavanj|objas.?nja)\w*)', r'(\s|^)((konfuz|\w*sigurn|supe|izuzetn|dobr|los.?|\w*brz|\w*spor|\w*obimn|zna|\w*razum|odlic.?n|monoton|zamaraju|interesant|\w*j

	asn|isto|razlic.?i|\w*glasn|\w*tih|\w*c.?est|\w*retk)\w*)[ ](\w+)?((predaj|objas.?xnjav|prelazi)\w*)', r'(\s|^)((predaj|objas.?xnjav)\w*)(\w+)?[ ](konfuz|\w*sigurn|supe|izuzetn|dobro|los.?|\w*brz|\w*spor|\w*ob imn|zna|\w*razum|odlic.?n|monoton|zamaraju|interesant|\w*jasn|is to|razlic.?i|\w*glasn|\w*tih|\w*c.?est|\w*retk)\w*)', r'([\n\r\s]|^)((svi|sve|svaka)[ ]pohval\w*)[ ](\w*)?(\w*)?((rad|izlaganj)\w*)']
Однос	[r'(\s|^)odnos sa (stud|prof)\w*', r'(\s|^)(odnos (sa|prema)?)((?:\S+){0,4})(stud|prof|asis|nastavnik)\w*', r'(\s|^)(interakci\w+) sa (stud|prof|asis|nastavnik)\w*', r'(\s|^)(interakci\w*|izmedju|interakci\w* izmedju)[ ](stud|uc.?eni)\w*)[ ](i)?((prof|asis|nastavni..)\w*)', r'(\s|^)(interakci\w*|izmedju|interakci\w* izmedju)[ ](prof|asis|nastavni..)\w*)[ ](i)?((stud|uc.?eni)\w*)', r'(\s|^)((ne\w+) po)((?:\S+){0,4}) (stud|uc.?eni)\w*', r'(\s|^)(posvec.?en\w*)[ ](stud|uc.?eni)\w*', r'(\s|^)(posvec.?uj\w*)[ ]((?:\S+){0,4})(stud|uc.?en)\w*', r'(\s|^)(sprem\w*)[ ](da)[ ]((pomogn|odgov.r|\w*savet|saslus.?)\w*)', r'(\s|^)(izadj\w*)[ ](u susret)', r'(\s|^)(((o|za)tvoren|\w*sprem|\w*zainteresov|\w*raspoloz.?e|tez.?a k|tes.?k)\w*)[ ](\w*)?(za saradnju)', r'(\s|^)((\w*razumev)\w*)[ ]((od|sa) stran)[ ](asis|prof|nastavn|predavac.?)\w*)']
Предмет	[r'(\s|^)(jedan od naj\w*)[ ]((predmeta|kurs)\w*)', r'([\n\r\s]|^)((svi|sve|svaka)[ ]pohval\w*)[ ](\w*)?(\w*)?((predme)\w*)']
Организација	[r'(\s|^)(mogl. bi|mislim da bi|trebalo bi|treba nam|potreb\w*|potrebno je|treb\w* bolje|sve pohvale|pohvale za|jedan od|bi trebalo|neophod\w*|izuzetno|svaka c.?ast)[ ](\w*)*?((naj|ne)?organiz)\w*', r'(\s|^)((re|ne|naj)?organiz)\w*[ ](\w*)*?(predmet|konsultacij|polaganj|nastav|vez.?b|predavanj|ispit|ocenji vanj|projek|teorij|kolokvij|seminar)\w*', r'(\s|^)(predmet|konsultacij|polaganj|nastav|vez.?b|predavanj|ispit|o cenjivanj|projek|teorij|kolokvij|seminar)\w*[ ](\w*)*?((re|ne|naj)?organiz)\w*', r'(\s|^)(\w*usklad|\w*usaglas|poveza)\w*[ ](\w*)*?(predmet|konsultacij|polaganj|nastav|vez.?b|predavanj|ispit|ocenji vanj|projek|teorij|kolokvij|seminar)\w*(i)(predmet|konsultacij|polaganj|nastav|vez.?b|predavanj|ispit|ocenjiva nj|projek|teorij|kolokvij|seminar)\w*', r'(\s|^)(predmet|konsultacij|polaganj|nastav|vez.?b|predavanj|ispit|o cenjivanj|projek|teorij|kolokvij|seminar)\w*(i)(predmet|konsultacij|polaganj|nastav|vez.?b|predavanj|ispit|ocenjiva

	nj|projek|teorij|kolokvij|seminar)\w*[ ](\w*)*?(\w*usklad|\w*usaglas|poveza)\w*', r'(\s|^)(predmet|konsultacij|polaganj|nastav|vez.?b|predavanj|ispit|ocenjivanj|projek|teorij|kolokvij|seminar)\w*(\w*usklad|\w*usaglas|poveza)\w*[ ](\w*)*?(predmet|konsultacij|polaganj|nastav|vez.?b|predavanj|ispit|ocenjivanj|projek|teorij|kolokvij|seminar)\w*', r'([\n\r\s]|^)((svi|sve|svaka)[ ]pohval\w*)[ ](\w*)?(\w*)?((organiz)\w*)']
Материјали	[r'(\s|^)(\w*vis.?|naprav|\w*postoj|mal|pun|slab|ima|nema|ponavlj|potreb|dob.?r|los.?|fal|\w*slagan|\w*podudar|zastare|\w*obimn|super|nedostaj)\w*[ ](?!((predaj|objas.?nja)\w*))(\w*)?(\w*)?(gradiv|materijal|slajd|literatu|skript|praktikum|prezentacij|udz.?benik|knjig|slik|skic)\w*', r'(\s|^)(?!((predaj|objas.?nja)\w*))(gradiv|materijal|slajd|literatu|skript|praktikum|prezentacij|udz.?benik|knjig|slik|skic)\w*[ ](?!((predaj|objas.?nja)\w*))(\w*)?(\w*vis.?|\w*postoj|mal|pun|slab|ima|nema|ponavlj|potreb|preobim|dob.?r|los.?|fal|\w*slagan|\w*podudar|zastare|\w*obimn|super)\w*', r'([\n\r\s]|^)((svi|sve|svaka)[ ]pohval\w*)[ ](\w*)?(\w*)?((gradiv|materijal|slajd|literatu|skript|praktikum|prezentacij|udz.?benik|knjig|slik|skic)\w*)']
Остало	[r'([\n\r]|^)((svi|sve)[ ])(\w*)?(supe|ok|najbolj|odlic.?n|dobr|los.?)\w*(\w*)?$', r'([\n\r]\|^)((svi\|sve)[ ])(\w*)?(supe\|ok\|najbolj\|odlic.?n\|dobr\|los.?)\w*[ ](?!(gradiv\|materijal\|slajd\|literatu\|skript\|praktikum\|prezentacij\|udz.?benik\|knjig\|slik\|skic\|nastav\|vez.?b\|predav\|asis\|prof\|nastavn)\w*)$']

“You think it's the end, But it's just the beginning.”
Bob Marley & The Wailers

Овај Образац чини саставни део докторске дисертације, односно докторског уметничког пројекта који се брани на Универзитету у Новом Саду. Попуњен Образац укоричити иза текста докторске дисертације, односно докторског уметничког пројекта.

План третмана података

Назив пројекта/истраживања
Аутоматско издвајање мишљења из текстуалних коментара студентских анкета
Назив институције/институција у оквиру којих се спроводи истраживање
а) **Универзитет у Новом Саду, Факултет техничких наука** б) в)
Назив програма у оквиру ког се реализује истраживање
Истраживање се реализује у оквиру израде докторске дисертације на студијском програму Рачунарство и аутоматика
1. Опис података
1.1 Врста студије *Укратко описати тип студије у оквиру које се подаци прикупљају* **Докторска дисертација** _____ 1.2 Врсте података **а) квантитативни** **б) квалитативни** 1.3. Начин прикупљања података **а) анкете, упитници, тестови** б) клиничке процене, медицински записи, електронски здравствени записи в) генотипови: навести врсту _____ **г)** административни подаци: навести врсту **Анонимизовани резултати анкета и статистички подаци**

д) узорци ткива: навести врсту_____

ђ) снимци, фотографије: навести врсту_____

е) текст, навести врсту **Литературни извори**

ж) мапа, навести врсту _____

з) остало: описати **Нумерички експерименти**

1.3 Формат података, употребљене скале, количина података

1.3.1 Употребљени софтвер и формат датотеке:

а) Excel фајл, датотека .xlsx

b) SPSS фајл, датотека _____

c) PDF фајл, датотека .pdf

d) Текст фајл, датотека .txt, .doc, .docx

е) JPG фајл, датотека _____

f) Остало, датотека _____

1.3.2. Број записа (код квантитативних података)

а) број варијабли **велики број**

б) број мерења (испитаника, процена, снимака и сл.) **велики број**

1.3.3. Поновљена мерења

а) да

б) не

Уколико је одговор да, одговорити на следећа питања:

а) временски размак између поновљених мера је _____

б) варијабле које се више пута мере односе се на _____

в) нове верзије фајлова који садрже поновљена мерења су именоване као _____

Напомене: _____

Да ли формати и софтвер омогућавају дељење и дугорочну валидност података?

а) Да

б) Не

Ако је одговор не, образложити **Софтвер је формиран на основу приватних података тако да није могуће дељење података али је могуће слободно коришћење софтвера.**

2. Прикупљање података

2.1 Методологија за прикупљање/генерисање података

2.1.1. У оквиру ког истраживачког нацрта су подаци прикупљени?

а) експеримент, навести тип **Нумерички експеримент**

б) корелационо истраживање, навести тип _____

ц) анализа текста, навести тип **Прикупљање података анализом доступне литературе**

д) остало, навести шта **Анонимизовани коментари из званичних анкета приватне природе**

2.1.2 Навести врсте мерних инструмената или стандарде података специфичних за одређену научну дисциплину (ако постоје).

_____

2.2 Квалитет података и стандарди

2.2.1. Третман недостајућих података

а) Да ли матрица садржи недостајуће податке? Да **Не**

Ако је одговор да, одговорити на следећа питања:

а) Колики је број недостајућих података? _____

б) Да ли се кориснику матрице препоручује замена недостајућих података? Да Не

в) Ако је одговор да, навести сугестије за третман замене недостајућих података

_____

2.2.2. На који начин је контролисан квалитет података? Описати

Квалитет података је контролисан поређењем експерименталних и теоријских података

_____

2.2.3. На који начин је извршена контрола уноса података у матрицу?

Контрола уноса података у матрицу је извршена од стране административног лица Факултета техничких наука која су уносила податке из анкета у електронски облик

_____

3. Третман података и пратећа документација

3.1. Третман и чување података

3.1.1. Подаци ће бити депоновани у _____ *репозиторијум.*

3.1.2. URL адреса _____

3.1.3. DOI

_____

3.1.4. Да ли ће подаци бити у отвореном приступу?

а) Да

б) Да, али после ембарга који ће трајати до _____

в) Не

Ако је одговор не, навести разлог

Постоји ограничење о приступу подацима од стране Факултета техничких наука, Универзитета у Новом Саду, као и ризик од злоупотребе, неовлашћеног преузимања, обраде и објављивања целине или дела прикупљених и обрађених података истраживања.

3.1.5. Подаци неће бити депоновани у репозиторијум, али ће бити чувани.

Образложење

Подаци неће бити у отвореном приступу. Подаци се чувају у електронској форми на рачунарима одговорних и овлашћених лица.

3.2 Метаподаци и документација података

3.2.1. Који стандард за метаподатке ће бити примењен?

Не примењује се стандард за метаподатке.

3.2.1. Навести метаподатке на основу којих су подаци депоновани у репозиторијум.

Не примењује се стандард за метаподатке.

Ако је потребно, навести методе које се користе за преузимање података, аналитичке и процедуралне информације, њихово кодирање, детаљне описе варијабли, записа итд.

3.3 Стратегија и стандарди за чување података

3.3.1. До ког периода ће подаци бити чувани у репозиторијуму?

_____

3.3.2. Да ли ће подаци бити депоновани под шифром? Да **Не**

3.3.3. Да ли ће шифра бити доступна одређеном кругу истраживача? Да **Не**

3.3.4. Да ли се подаци морају уклонити из отвореног приступа после извесног времена?

Да **Не**

Образложити

---

4. Безбедност података и заштита поверљивих информација

Овај одељак МОРА бити попуњен ако ваши подаци укључују личне податке који се односе на учеснике у истраживању. За друга истраживања треба такође размотрити заштиту и сигурност података.

4.1 Формални стандарди за сигурност информација/података

Истраживачи који спроводе испитивања с људима морају да се придржавају Закона о заштити података о личности (*https://www.paragraf.rs/propisi/zakon_o_zastiti_podataka_o_licnosti.html*) и одговарајућег институционалног кодекса о академском интегритету.

4.1.2. Да ли је истраживање одобрено од стране етичке комисије? Да **Не**

Ако је одговор Да, навести датум и назив етичке комисије која је одобрила истраживање

_____

4.1.2. Да ли подаци укључују личне податке учесника у истраживању? **Да** Не

Ако је одговор да, наведите на који начин сте осигурали поверљивост и сигурност информација везаних за испитанике:

а) Подаци нису у отвореном приступу

б) Подаци су анонимизирани

ц) Остало, навести шта

_____

_____

5. Доступност података

5.1. Подаци ће бити

а) јавно доступни

б) доступни само уском кругу истраживача у одређеној научној области

ц) затворени

Ако су подаци доступни само уском кругу истраживача, навести под којим условима могу да их користе:

_____

Ако су подаци доступни само уском кругу истраживача, навести на који начин могу приступити подацима:

5.4. Навести лиценцу под којом ће прикупљени подаци бити архивирани.

Ауторство–некомерцијално–без прераде

6. Улоге и одговорност

6.1. Навести име и презиме и мејл адресу власника (аутора) података

Никола Николић, мејл адреса: nikola.nikolic@uns.ac.rs

6.2. Навести име и презиме и мејл адресу особе која одржава матрицу с подацима

Никола Николић, мејл адреса: nikola.nikolic@uns.ac.rs

6.3. Навести име и презиме и мејл адресу особе која омогућује приступ подацима другим истраживачима

Напомена: Подаци нису доступни.